



人工智能治理白皮书

中国信息通信研究院

中国人工智能产业发展联盟

2020年9月

版权声明

前　言

当前，全球科技革命和产业变革孕育兴起，以深度学习、跨界融合、人机协同、群智开放、自主操控为特征的新一代人工智能不断取得突破，已成为新一轮科技革命和产业变革的重要驱动力量。在极大提升人类生产生活品质的同时，新一代人工智能具有的通用目的性、算法黑箱性以及数据依赖性等技术特性，引发了社会、企业、个人等不同维度的风险和挑战，从而对治理提出了专业化、多元化、敏捷化、全球化的迫切需要。因此，人工智能治理这一概念应运而生，成为国际社会广泛关注和研究的方向。

人工智能治理是一项复杂的系统工程，既需要明确治理原则、目标以及厘清治理主体，又需要提出切实有效的治理措施。全球正在逐步构建起人工智能治理框架，以坚持科技造福人类，平衡创新发展与有效治理的关系作为治理目标，采用多元主体参与、协同共治的治理机制，通过制定伦理原则、设计技术标准、确立法律法规等综合治理手段，推动人工智能健康有序发展。

人工智能的治理体系需要“柔性的伦理”和“硬性的法律”的共同构建。一方面，以伦理为导向的社会规范体系，可以为人工智能技术层面的开发和应用提供价值判断标准，约束和指导各方对人工智能进行协同治理。众多国际组织、国家和企业选择从伦理角度入手，试图确立人工智能的基本伦理规范，探索清晰的道德边界，并积极构建人工智能伦理的落地机制和体系。另一方面，以法律为保障的风险防控体系，依靠国家强制力划定底线，可以防范和应对人工智能技术带来的诸多风险。整体来看，人工智能相关立法正从慌乱走向理性，从源头治理走向综合治理，从粗放治理走向精细治理。同时，在自动驾驶、深度伪造、智能金融、智能医疗等场景下，人工智能立法取得率先突破，积累了一定的监管经验。

本白皮书从人工智能第三次浪潮开辟的治理新需要出发，结合当前人工智能发展阶段，分析了全球目前的人工智能治理机制，重点介绍了两项主要的治理举措，即以伦理为导向的社会约束体系和以法律为保障的风险防控体系，最后对人工智能治理的未来发展方向进行了展望，期待为社会各方提供有益参考。

目 录



图 目 录



表 目 录

一、人工智能开辟治理新需要

人工智能是指，用机器模拟、实现或延伸人类的感知、思考、行动等智力与行为能力的科学与技术。作为引领未来的战略性技术，人工智能是新一轮科技革命和产业变革的重要驱动力量，已经成为国际竞争的新焦点、经济发展的新引擎。人工智能在极大提升人类生产水平和生活品质的同时，也带来了新风险、引发了新问题，对一个国家治理能力的法治化、智能化、专业化水平提出全新要求。目前，国际社会纷纷呼吁加强人工智能治理，发展安全可信、负责任的人工智能。

（一）人工智能时代的兴起

人工智能概念的提出，始于 1956 年美国达特茅斯会议。人工智能至今已有 60 多年的发展历程，从诞生至今经历了三次发展浪潮。前两次浪潮中，由于算法的阶段性突破而达到高潮，之后又由于理论方法缺陷、产业基础不足、场景应用受限等原因而没有达到人们最初的预期，并导致了政策支持和社会资本投入的大幅缩减，从而两次从高潮陷入低谷。近年来，在移动互联网、大数据、超级计算、传感网、脑科学等新理论新技术以及经济社会发展强烈需求的共同驱动下，以

深度学习、跨界融合、人机协同、群智开放、自主操控为特征的新一代人工智能技术不断取得新突破，迎来了人工智能的第三次发展浪潮。

1. 人工智能驱动新一轮科技革命

人工智能是当前科技革命的制高点，以智能化的方式广泛联结各领域知识与技术能力，释放科技革命和产业变革积蓄的巨大能量，成为全球科技战的争夺焦点。世界主要发达国家纷纷把发展人工智能作为提升国家竞争力的主要抓手，努力在新一轮国际科技竞争中掌握主导权，围绕基础研发、资源开放、人才培养、公司合作等方面强化部署。例如，美国为人工智能研发投入大量资金，确保其人工智能在全球的领先地位；英国利用其在计算技术领域的积累，致力于建设世界级人工智能创新中心；日本以建设超智能社会 5.0 为引领，旨在强化其在汽车、机器人等领域全球领先优势。

2. 人工智能打造经济发展新引擎

当前，以智能家居、智能网联汽车、智能机器人等为代表的人工智能新兴产业加速发展，经济规模不断扩大，正成为带动经济增长的重要引擎。普华永道提出，人工智能将显著提升全球经济，到 2030

年，人工智能将促使全球生产总值增长 14%，为世界经济贡献 15.7 万亿美元产值。一方面，人工智能驱动产业智能化变革，在数字化、网络化基础上，重塑生产组织方式，优化产业结构，促进传统领域智能化变革，引领产业向价值链高端迈进，全面提升经济发展质量和效益。另一方面，人工智能的普及将推动多行业的创新，大幅提升现有劳动生产率，开辟崭新的经济增长空间。据埃森哲预测，2035 年人工智能将推动我国劳动生产率提高 27%，经济总增加值提升 7.1 万亿美元。

3. 人工智能显著提升社会生活质量

人工智能在教育、医疗、养老等民生服务领域应用广泛，推动服务模式不断创新，服务产品日益优化，创新型智能服务体系逐步形成。在医疗方面，人工智能不断提升医疗水平，特别是在新冠肺炎疫情期间，人工智能在疫情监测、疾病诊断、药物研发等方面发挥了重要作用。在教育方面，人工智能的应用加快了开放灵活的教育体系的建设工作，能够实现因材施教，推动个性化教育发展，进一步促进教育公平和提升教育质量。在养老方面，人工智能在助残养老领域的应用不

断丰富和创新，在帮助残疾人和老人提升生活自理能力和尊严感方面发挥重要作用。例如，护理型机器人通过与照护对象进行交互性治疗，可以降低老年人的孤独感，极大改善老年人的生活。

（二）人工智能的技术特性

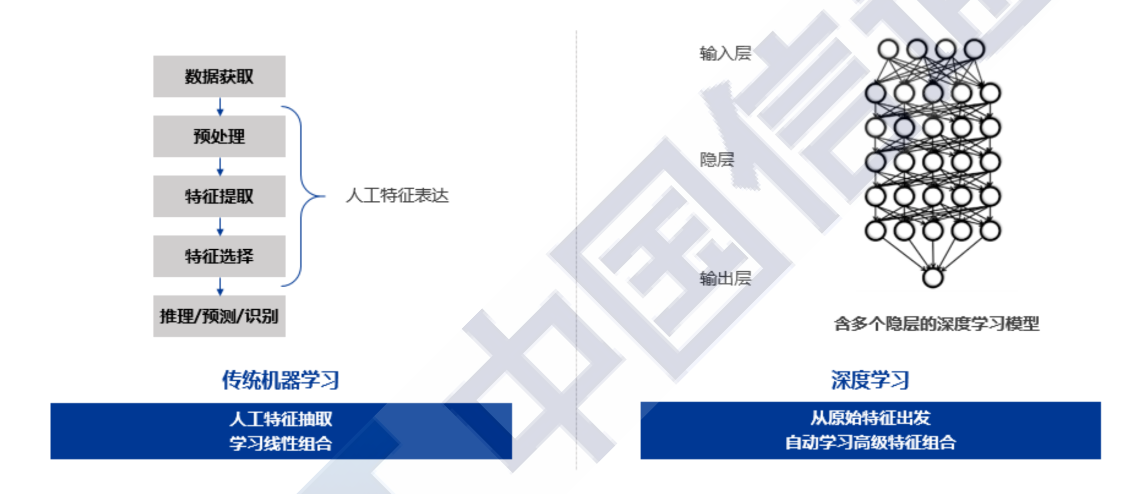
人类正处于一个前所未有的时期，互联网、移动通信、物联网、云计算、大数据、人工智能等信息技术正在推动社会逐步进入智能化时代。相较于其他领域的新技术，由于新一代人工智能具有的技术特性，使其更为复杂、更不可控、更难预测。

1. 通用目的性，导致风险更为普遍

人工智能作为一项通用目的技术，可以应用到自动驾驶、智能制造、智慧城市、智慧医疗等诸多领域场景中，并且能够与大数据、云计算等数字技术互补互促使用，有着极强的技术溢出效应，对经济社会高质量发展显现出强劲的引领带动作用。但与此同时，人工智能技术风险发生的范围会随着应用场景的日趋广泛而逐步扩大，问题发生的可能性也会随着其应用频次的增长而持续提高，若不及时有效治理，将会严重影响人类生产生活。

2. 算法黑箱性，导致过程更难解释

目前，以深度学习算法为核心的人工智能算法模型被普遍应用，但由于其算法结构中存在多个“隐层”，导致输入数据和输出结果之间的因果逻辑关系难以清楚解释，用户只能被动接受由算法带来的结果而无法洞悉其运行过程，从而形成一种技术“黑箱”。此外，人工智能算法模型还具有自适应、自学习等特性，导致其极易偏离人类预设的目标，其复杂程度愈发超出人类理解范畴。



来源：资料整理

图 1　深度学习与传统机器学习的差别图

3. 数据依赖性，导致结果更不可控

数据是人工智能模型训练及优化的“燃料”，是人工智能算法做出正确、公平、合理决策的基础保障。输入数据的数量规模、准确性、

通用性、包容性、全面性等质量因素将直接决定训练得到的模型的质量。同时，数据本质上是社会价值观的缩影与映射，因此也会包含一些落后的价值观与社会偏见。若未能对数据质量进行有效把控，人工智能算法模型便很可能习得数据中的偏见谬误，并将其反映到训练结果中，致使人工智能系统的功能行为及其影响变得更不可控。

（三）人工智能引发不同维度的风险

作为引领未来的战略性技术，人工智能在打造经济发展新引擎、推动人类文明迈上新台阶的同时，也给社会、企业和个人带来了不同维度的风险挑战。

1. 社会维度：影响社会稳定

一是冲击就业格局，加剧财富分化。智能的算法、机器对传统人工的替代在解放人力劳动者的同时，直接带来了对就业的冲击。从事重复性、机械性等工作的劳动者更容易被人工智能替代工作。据麦肯锡报告推测，到 2030 年机器人将取代 8 亿人的工作。与历史上的技术革命类似，人工智能的发展同样会导致利益的分化与重构，新创造的社会财富将会不成比例地向资本一方倾斜，低收入与受教育程度较

低的人群将在新一轮的社会资源分配中处于严重的不利地位。

二是影响政治进程，抹黑政治人物。人工智能在社交服务中的应用能够影响政治进程，利用机器人水军可以进行舆论干预。例如，剑桥分析公司利用人工智能，辅助进行竞选策略，影响美国大选结果。此外，可以利用深度伪造等智能信息服务技术制作关于政治人物的虚假负面视频。例如，2018 年 4 月，美国前总统奥巴马的脸被“借用”来攻击特朗普总统，视频在 YouTube 上被转发 500 多万次。

三是侵害事件频发，危及公共安全。人工智能安全事故、侵害事件频发，引发社会各界普遍关注。例如，特斯拉 Model S 在美国和我国境内都曾发生过自动驾驶致死事故和数起交通事故；2018 年委内瑞拉总统在公开活动中受到无人机炸弹袭击，这是全球首例利用人工智能产品进行的恐怖活动；2018 年，优步的一辆自动驾驶测试车在进行路试时发生事故，导致一行人死亡。

2. 企业维度：加大合规难度

一是不良信息频现，企业审核能力不足。如果向人工智能系统输入不完整、不正确、质量不高的数据，则会产生不良或者歧视性信息，

即所谓的“垃圾进，垃圾出”。例如微软公司的人工智能聊天机器人Tay 上线后，被网民“教坏”，发布诽谤性的、歧视性的推文。此外，人工智能技术极大地促进数字内容产业的繁荣。预计到 2022 年，全球互联网流量将达到每秒 7.2 PB。企业试图依靠传统审核模式实现内容的准确判断并及时应对信息爆炸引发等各类问题，越发捉襟见肘。

二是法律责任不明，陷入责任划分困境。由于当前人工智能产品在问题回溯上存在不可解释环节，而且现行立法也未明确界定人工智能的设计、生产、销售、使用等环节的各方主体责任与义务，这给人工智能安全事件的责任认定和划分带来严峻挑战。例如，人工智能医疗助理（例如 IBM 的“沃森医生”）给出危险错误的癌症医疗建议时的责任认定等问题。而且，当人工智能出现自主决策能力后，自动驾驶汽车因独立智能决策致损时，如何确定侵权主体及划分责任。

三是知识产权保护不足，版权认定困难。目前，各国就人工智能生成物所包含的权利类型和权利归属存有争议，人工智能创作物的版权保护仍普遍面临法律滞后问题。例如，澳大利亚法院判定，利用人工智能生成的作品不能由版权保护，因为它不是人类制作的。如果人

工智能创作物得不到法律有力的保护，会使得人工智能生成信息的复制和扩散门槛更低，影响投资人、创造人投入人工智能创作的积极性。

3. 个人维度：侵犯基本权益

一是算法偏见现象，影响公平正义。人工智能算法并非绝对客观世界的产物，算法偏见不仅是技术问题，更涉及到对算法处理的数据集质量的完整性、算法设计者的主观情感偏向、人类社会所固有的偏见、甚至不同地区文化差异等各方面问题。例如，美国一些法院适用的风险评估算法 COMPAS 被发现对黑人造成了系统性歧视。人脸识别软件 Rekognition，曾将美国国会议员中的 28 人误判为罪犯。

二是信息收集多样，侵犯个人隐私。随着人脸识别、虹膜识别等应用的普及，人工智能正在大规模、不间断地收集、使用敏感个人信息，个人隐私泄露风险加大。例如，变脸应用“ZAO”因用户协议过度攫取用户授权、存在数据泄漏问题而被监管机构约谈要求自查整改。杭州一动物园因启用人脸识别技术，强制收集游客敏感个人信息而被诉至法院，成为我国“人脸识别第一案”。

三是滥用智能产品，侵犯人格尊严。利用深度伪造技术能实现将

人脸移转到色情明星的身体，伪造逼真的色情场景，使污名化他人及色情报复成为可能。例如，通过 DeepNude 软件，输入一张完整的女性图片就可一键生成相应的裸照。此外，还发生过亚马逊智能音箱劝主人自杀等事件。

（四）人工智能治理的新挑战

人工智能技术的发展已经进入了某些科技领域的“无人区”，人工智能除技术本身可能发生问题之外，诸多应用在使用过程中也存在负效应。但目前，相应的风险防控机制和规则制定相对滞后，不可控的预期与担忧使得人工智能在创新上面临巨大的压力，人工智能治理也就成为了人工智能技术和应用发展到一定阶段的必然结果。联合国全球治理委员会对治理的概念进行了界定，认为“治理”是指“各种公共的或私人的个人和机构管理其共同事务的诸多方法的总和，是将相互冲突的或不同利益得以调和，并采取联合行动的持续过程”。当前，虽然各界对于人工智能治理的内涵没有统一的界定，但本白皮书将人工智能治理解释为，国际组织、国家、行业组织、企业等多主体对人工智能研究、开发、生产和应用中出现的公共安全、道德伦理等

问题进行协调、处理、监管和规范的过程。人工智能治理既需合理利用人工智能的优势，又要善于规避人工智能的负效应，以推动全人类社会福祉。

1. 治理专业化的需要

面对人工智能可能引发的新的全球焦虑，人工智能监管机构应及时察觉、管控危机、防范潜在风险。但目前，各国人工智能监管机构大多由政治实体组成，在应对人工智能等新议题的挑战时，需要与专业知识共同体形成有效联动。因此，监管机构应当以深厚的专业化知识作为治理基石，构建起涵盖人工智能相关的数学、计算机科学等学科的专业化治理团队，为监管提供必要的知识储备和智力支持。

2. 治理多元化的需要

人工智能本身是一门综合性的前沿学科和高度交叉的复合型学科，研究范畴广泛而又异常复杂。人工智能的治理需要计算机科学、数学、认知科学、神经科学和社会科学等学科深度融合，这引发了对于人工智能治理主体多元构成的需求。近年来，国际组织、各国政府、行业组织和企业等各类主体也在积极探索多元主体参与的协同共治

治理格局。因为只有在治理过程中，不断推动多领域间进行广泛交流和合作，采用灵活多元的方式，才能避免某些不可逆的问题出现。

3. 治理敏捷化的需要

人工智能的产业革命呈现出高速迭代化的特征，各种细分领域的产业化应用层出不穷。新兴业态呼唤新的治理方案，在治理原则、治理主体和治理手段上有别于传统治理框架，引发了对于人工智能治理敏捷化的需要。在治理过程中，需要通过不断提升技术手段，优化治理机制，及时发现和解决可能引发的风险，对更高级人工智能的潜在风险持续开展研究和预判，确保人工智能朝着有利于社会的方向发展。

4. 治理全球化的需要

人工智能作为一种通用性技术，所引发的人员失业、公共安全等问题具有全球性，需要全球共同面对并达成全球性共识，而且人工智能的不稳定与多样化应用又呼唤相对统一的治理规则与国际合作。只有世界各国共同寻求有效路径，探索全球问题的解决之道，才能使人工智能在可见的甚或遥远的未来更好地造福于人类，及时管控可能的危机以及防范潜在风险。



二、全球构建人工智能治理机制

人工智能治理是一项复杂的系统工程，既需要明确治理原则及目标、厘清治理主体，又需要提出切实有效的治理措施。为此，人工智能治理应当构建由政府、行业组织、企业以及公众等多元主体共同参与、协同合作的多层次的治理体系，通过制定伦理原则、设计技术标准、确立法律法规等多种举措，实现科技向善、造福人类的总体目标愿景，推动人工智能健康有序发展。



来源：资料整理

图 2　人工智能治理机制图

（一）治理目标：坚持科技造福人类，平衡创新发展与有效治理

总体来说，人工智能治理应以科技造福人类为总体目标，既要不断释放人工智能所带来的技术红利，也要精准防范并积极应对人工智能可能带来的风险，需要平衡好人工智能创新发展与有效治理的关系，持续提升有关算法规则、数据使用、安全保障等方面的治理能力，为人工智能营造规范有序的发展环境。

1. 不断释放人工智能带来的技术红利和价值

人工智能作为一项新型通用目的性技术，将在改造升级传统产业、培育新兴产业、加速实体经济转型、保障改善民生等经济社会发展诸多关键环节发挥重要作用。例如，阿根廷、巴西等国在其发布的人工智能国家战略中明确指出“人工智能应提升各利益攸关方的福祉”“将人工智能对科技发展、竞争力与生产力提升、公共福利增强的促进作用最大化”，最大程度释放人工智能技术红利。在我国，目前人工智能正在与制造业、交通、物流、农业、公共服务等攸关社会民生的行业进行不同程度的融合，逐渐创造出了新产品、新服务、新业态、新模式，形成创新驱动力、促进传统行业变革、促进实体经济增长，不断释放人工智能技术红利和潜在价值，推动实现经济高质量发展。

2. 精准防范并应对人工智能可能带来的风险

人工智能在为人类社会发展带来更多便利、更高效率的同时，也会进一步模糊机器世界与人类世界的边界，导致诸如算法歧视、隐私保护、权利保障等风险问题，甚至会引发社会失业、威胁国家安全等严峻挑战。鉴于人工智能带来的风险涉及范围广、影响大，因此有必要从全球治理的高度，重新审视并思考如何精准防范并有效应对解决人工智能所带来的风险挑战，以避免人工智能对人类社会发展产生的负面影响。例如，韩国在《人工智能国家战略》中提出“防止人工智能产生负面效应，制定人工智能伦理体系，推进监测人工智能信赖度、安全性等的质量管理体系建设”等引导人工智能安全发展的要求。

平衡好人工智能创新发展与有效治理的关系是关键。一方面，过于严苛的治理方式会限制人工智能技术的创新与进步，导致任何技术创新都步履维艰。但另一方面，没有任何监管与规制的人工智能极易“走偏”甚至“误入歧途”，给人类社会带来风险与危害，违背科技造福人类的目标愿景。因此，应当找到创新发展与有效治理之间的平衡点，坚持安全可控的治理机制，将开放创新的技术发展并重，给予

技术进步与市场创新适当的试错、调整空间，对人工智能发展既不简单粗暴、一刀切似地扼杀，也不任其自由泛滥，而是要充分发挥出多元主体协同共治的效能，使各方各司其职、各尽其力，把握好治理原则，守住治理底线，确保人工智能产业创新活力与发展动力，提升公众在使用人工智能技术及产品时的获得感、安全感、幸福感。

（二）治理模式：打造多元主体参与、协同共治的模式

人工智能治理的重要特征之一是治理主体的多元化，其依赖于包括国家政府、行业组织、企业、公众在内的各利益攸关方的参与合作、各司其职、各尽其能，以适当的角色、最佳的方式协同共治，从而构建严谨、全面、有效的全新治理模式。

1. 国家政府是执行治理规则的核心主体

国家政府在人工智能治理中发挥着领导性作用，主要体现在国家层面上统领人工智能技术研发与治理框架的搭建，专业管理机构的设立，以及政策与法律规则的制定等方面。国家政府作为肩负公共事务管理职责的公权力机关，是公共利益和广大民意的代言人，同时也是国家安全和社会稳定的捍卫者，对人工智能技术进行治理便是应有之

义。为此，各国政府增设专业管理机构，积极布局人工智能技术的研发与投资路线，监督和管理为人工智能发展设定的标准和规则。美国成立的管理机构较多，分别聚焦于军事、技术研发等方面的监督管理。欧盟以及英国则更加关注经济、产业方面的监管议题，并注重数据治理与保护。日本重点关注智能制造、健康医疗、智能交通等领域的应用与监管。整体上来看，各国管理机构的人员背景具有多元化，机构职能具有多样化的特点。

表 1 各国人工智能管理机构

国家	监管机构名称	监管相关职责	建立时间	负责部门
美国	机器学习与人工智能分委会	监督各行业、研究机构以及政府部门的人工智能研发	2016 年 5 月	美国国家科学与技术委员会
	人工智能专门委员会	负责审查联邦机构的人工智能领域投资和开发方面的优先事项	2018 年 5 月	白宫科技政策办公室、美国国家科学与技术委员会、国防部高级研究计划局等
	联合人工智能中心	监管国防机构人工智能工作	2018 年 6 月	美国国防部
	人工智能国家安全委员会	考察并监督人工智能技术应用在军事中的情况，评估其安全、伦理、对国际法影响等风	2018 年 11 月	美国众议院武装部队新型威胁与能力小组委员会

国家	监管机构名称	监管相关职责	建立时间	负责部门
		险		
欧盟	人工智能高级小组	研究并起草人工智能监管框架、并指导欧洲相关企业进行落实	2018 年 6 月	欧盟委员会
英国	人工智能理事会	监督英国人工智能战略实施并为政府提出建议	2018 年 4 月	英国政府人工智能办公室
	数据伦理和创新中心	审查、监管现有的数据（包括人工智能）治理格局，并就其安全、道德、创新使用为政府提出建议	2018 年 11 月	英国政府
	人工智能特别委员会	负责提供人工智能发展建议，为 AI 发展设定标准和规则	2018 年 4 月	英国上议院
法国	人工智能伦理委员会	监督军用人工智能的发展	2019 年 4 月	法国政府
日本	人工智能技术战略会议	国家层面的人工智能综合管理机构，负责政策及应用的监管	2016 年 4 月	日本政府
印度	人工智能伦理委员会联盟	制定人工智能产品研发标准	2018 年 6 月	印度政府
墨西哥	人工智能办公室	规范人工智能健康发展	2018 年 6 月	墨西哥政府
中国	新一代人工智能发展规划推进办公室	研究人工智能相关法律、伦理、标准、社会问题以及治理议题	2017 年 11 月	国家科技体制改革和创新体系建设领导小组

来源：资料整理

2. 政府间国际组织是引导治理方向的重要力量

人工智能的发展具有跨国界、国际分工的特征，需要政府间国际组织加强国家间协调合作。因此，政府间国际组织在全球人工智能治理中扮演着引导者和推动者的角色。首先，政府间国际组织引导人工智能领域形成大国共识。由于各国在人工智能技术研发的关注与投入不同，关于人工智能治理的规则率先在发达国家形成和扩散。政府间国际组织作为引领国际规则制定的风向标，也在早期就针对人工智能与监管展开讨论，汲取各国关于人工智能治理的原则性宣言，引导人工智能治理稳步迈向并达成国际共识。其次，政府间国际组织推动人工智能治理规则的全球共享。人工智能技术在各国发展参差不齐，多数发展中国家和不发达国家还并未将人工智能治理纳入国家战略。由此，政府间国际组织前瞻性地发布人工智能治理规则，推动缩短国家间数字鸿沟，促进世界各国人工智能技术的协调、健康、共享发展。

3. 行业组织是协调多方治理的积极推动者

行业组织作为兼顾服务、沟通、自律、协调等功能的社会团体，是协调人工智能治理、制定人工智能产业标准的先行者和积极实践者。

其中，行业组织包括行业协会、标准化组织、产业联盟等机构，代表性的行业协会包括电气与电子工程师协会（IEEE）、美国计算机协会（ACM）、人工智能促进协会（AAAI）等，标准化组织包括国际标准化组织（ISO）、国际电工委员会（IEC）等。产业联盟包括国际网络联盟、我国的人工智能产业技术创新战略联盟、人工智能产业发展联盟等。为推动行业各方落实人工智能治理要求，行业组织较早地开展了人工智能治理相关研究，积极制定人工智能技术及产品标准，并持续贡献着治理智慧。

表 2 全球人工智能标准文件

文件名称	核心内容	发布机构
《ISO/IEC 20005》（2013 年）	传感器网络标准化：术语与词汇、智能传感网络协同信息处理支持服务和接口	国际标准化组织、国际电工委员会
《ISO/IEC 30122》（2016 年）	人机交互标准化:框架与通用指南、构建与测试、翻译与本地化、语音命令注册管理	
《ISO/IEC 19944》（2017 年）	云计算标准化：互操作与可移植、跨设备数据与云服务的数据流	
《算法透明和可责性声明》（2017 年）	充分认识算法歧视、明确数据来源、提高可解释性与可审查性、建立严格的验证测试机制	美国计算机协会
《IEEE P7000》（2016 年）	系统设计期间解决伦理问题的模型过程的标准	电气和电子工程师协会

文件名称	核心内容	发布机构
《IEEE P7001》（2016 年）	自主系统的透明度的标准	
《IEEE P7002》（2016 年）	数据隐私处理的标准	
《IEEE P7003》（2017 年）	算法偏差注意事项	
《IEEE P7006》（2017 年）	个人数据人工智能代理标准	
《中文语音识别系统通用技术规范》《中文语音合成系统通用技术规范》《自动声纹识别（说话人识别）技术规范》《中文语音识别互联网服务接口规范》《中文语音合成互联网服务接口规范》	语音交互系列标准	全国信息技术标准化技术委员会
《共享学习系统技术要求》（2020 年）	共享学习的技术框架及流程、技术特性、安全要求	中国人工智能产业发展联盟

来源：资料整理

4. 企业是践行行业自律自治的中坚力量

企业在推动人工智能治理规则和标准落地上发挥着决定性作用，是践行治理规则和行业标准的中坚力量。企业作为人工智能技术的主要开发者和拥有者，掌握了资金、技术、人才、市场、政策扶持等大量资源，理应承担相关社会责任，严格遵守科技伦理、技术标准以及

法律法规，以高标准进行自我约束与监督，实现有效的行业自律自治。面对人工智能所引发的社会担忧与质疑，一些行业巨头企业也开始研究人工智能对社会经济、伦理等问题的影响，并积极采取措施确保人工智能可以造福人类。由此，企业是必不可少地人工智能治理规则践行者，也是确保人工智能技术向正确道路发展的重要防线。

5. 公众是监督治理效果的重要参与者

公众是人工智能技术、产品的主要服务对象，拥有对人工智能治理相关内容的监督、讨论、意见反馈等权利。因此公众要积极参与到治理规则制定中，适当介入相关监督、监管过程中，为人工智能治理献计献策，形成自下而上的协同治理模式，使人工智能服务真正地“以人为本，造福人类”。治理好人工智能，需要着力畅通各个治理主体间沟通渠道，加强多元主体间的对话与协商，合作制定为应对人工智能风险挑战的整体解决方案。因此，公众的参与是实现人工智能有效治理不可缺少的重要力量。

（三）治理手段：实施多层次、多样化的治理举措

人工智能治理手段主要包括伦理约束、技术应对、规范立法等三

个方面。

1. 伦理约束

现行的伦理制度是以规范人与人之间的关系和行为为主体的制度，通过社会普遍价值观来实现行为约束功能，是人类价值判断和行为取向的根由。然而，人工智能模拟“人的智力”的能力实现技术与人的“耦合”，作为相对主体直接“参与”“人—人”的传统道德关系，形成“人—人工智能”相互间的道德关系。因此，作为人工智能治理中极为重要的一环—人工智能伦理，也就成为了人工智能时代的必然产物。具体而言，人工智能伦理是指处理机器与人、机器和社会相互关系时应遵循的道理和准则，它既包括对技术本身的研究，也包括在符合人类价值的前提下对人机之间的关系研究。人工智能最大的弱点是缺乏直接的感受能力，在价值判断上存在弱点，而人工智能伦理准则所倡导的造福人类、避免伤害、公平正义等价值理念，为人工智能技术层面的开发和运用提供了价值判断标准。人工智能伦理不仅弥补了法律的空白，还作为试验性规范为立法的创制积累经验。而且，伦理由于脱离强制性立法规定的范畴，更易引发国际共同体成员在人工

智能伦理领域的探讨，也更有助于在全球形成人工智能伦理共识。

2. 技术应对

随着人工智能的持续发展，人工智能的技术治理方式也在不断突破创新，方法种类不断丰富，作用效果逐步提升。同时，人工智能技术本身也作为一项精准、高效、智能的技术治理工具，正在被尝试用来解决人工智能所带来的风险问题，应对未来智能社会中的安全挑战。具体来看，现阶段部分领先企业已通过数据筛选、算法设计、模型优化等技术手段，将伦理原则“嵌入”人工智能应用与产品中，从而解决诸如隐私泄露、算法偏见、非法内容审核等问题，以达到科技向善的目的。例如微软利用“单词嵌入”的自然语言处理工具解决文本搜索中的性别偏见问题。IBM 通过其自研的“AI Fairness 360”工具包，监测并报告算法在机器学习训练中可能产生的偏见或歧视，并降低其后续发生概率。Facebook 利用 eGLYPH、DeepText 等工具，审核并发现存在极端主义或涉及种族歧视等言论的新闻内容，并第一时间对其删除阻断。

3. 规范立法

立法是治理的重要手段，对于人工智能的治理活动，立法是其重要途径。对于人工智能的产业促进，破除人工智能发展的现实阻碍，规范立法是其重要手段，例如人工智能产品带来的责任归属问题，智能产品侵权等相关法律问题，都是对人工智能的发展促进过程中必须关注和解决的重要问题。针对人工智能的规范立法，我们应当考虑采取包容审慎、灵活弹性的规制方式。一方面要避免草率立法对人工智能发展带来的阻碍，另一方面也要跟上技术发展节奏，敏捷灵活对其进行规范规制。因此，一种思路是在人工智能发展初期先通过制定行业公约、伦理规范、技术指南等方式对其进行敏捷灵活的治理，待其发展相对成熟，便可出来相关严格法律对其进行约束管控。目前全球各国也都在陆续开展对人工智能相关应用场景规范立法工作。

三、以伦理为导向的社会规范体系

由于新一代人工智能具有的通用目的性、数据依赖性和算法黑箱性等特性，因此人工智能治理规则不能完全依赖有强制约束力的法律，而是需要“伦理”和“法律”的共同构建。人工智能伦理对于促进人工智能进步、防止人工智能异化具有重要作用，具有良好伦理素养的研发者和应用者，能够将自己的行为与造福人类结合起来，慎重对待影响人的生存发展的人工智能研发和使用，防范人工智能风险事故的发生。因此，人工智能伦理能够以较为柔和的形式引导、规范行业行为，最终实现“柔性治理”的目的。

目前，人工智能治理的首要问题就是形成一套人工智能伦理体系，借用这套伦理体系去约束和指导各方对人工智能进行协同治理。众多国际组织、国家和企业选择从伦理角度入手，试图确立人工智能的基本伦理规范，探索清晰的道德边界，构建人工智能伦理的落地机制和体系，以引导人工智能创新，寻求创新与风险的平衡。从整体上看，以制定主体为标准进行划分，全球范围内的人工智能伦理文件主要分为三类：国际组织文件、各国政府文件和产业界文件，包括宣言、原

则、计划、指南等多种类型。虽然各自名称不同，但内容相似度较高，并以软性规制为主，普遍关注增进人类福祉、技术包容公平、维护人类尊严、保障安全和隐私保护等伦理内容。[1]但是，这三类文件的关注点略有不同，国际层面的伦理文件在积极探索共识性伦理原则；各国层面的伦理文件主要服务于国家人工智能产业发展路径；企业层面的伦理文件更加关注如何将伦理理念践行于具体的产品和服务中。



来源：资料整理

图 3　人工智能伦理文件类型图

（一）国际层面：开始制定全球性伦理倡议

由于人工智能引发的风险具有全球性特征，因此对于人工智能需以全球为规范面开展治理。而由于全球各区域针对人工智能呈现出的价值观念、规范方式及约束路径并不相同，因此若要在全球范围内对

[1] 张浩,黄克同.迈向制度化的人工智能伦理[J].人工智能,2019(04):32-38.

人工智能问题进行规范，重点在于达成全球性人工智能治理方案，而由于治理方案形成的出发点在于形成内涵固定的价值框架。因此如何将作为价值观念根本性元素的伦理观念达成共识是塑造人工智能伦理价值体系的基础性工作。

各个国际组织纷纷提出人工智能的伦理要求，对人工智能技术本身以及其应用进行规制。这些人工智能的治理文件，都表现出各主体对人工智能技术发展的担忧——要利用人工智能技术实现生产效率的提高和社会的进步，这一切都要建立在对风险的了解和预防的基础上。目前，联合国秉持着国际人道主义原则，在 2018 年提出了“对致命自主武器系统进行有意义的人类控制原则”，还提出了“凡是能够脱离人类控制的致命自主武器系统都应被禁止”的倡议，而且在海牙建立了一个专门的研究机构（犯罪和司法研究所），主要用来研究机器人和人工智能治理的问题。二十国集团（G20）于 2019 年 6 月发布《G20 人工智能原则》，倡导以人类为中心、以负责任的态度开发人工智能，并提出“投资于 AI 的研究与开发、为 AI 培养数字生态系统、为 AI 创造有利的政策环境、培养人的能力和为劳动力市场转

型做准备、实现可信赖 AI 的国际合作”等具体细则。经济合作与发展组织（OECD） 2019 年 5 月发布《关于人工智能的政府间政策指导方针》,倡导通过促进人工智能包容性增长，可持续发展和福祉使人民和地球受益，提出了“人工智能系统的设计应尊重法治、人权、民主价值观和多样性，并应包括适当的保障措施，以确保公平和公正的社会”的伦理准则。国际电气和电子工程师协会（IEEE）2019 年发布《人工智能设计伦理准则》（正式版），通过伦理学研究和设计方法论，倡导人工智能领域的人权、福祉、数据自主性、有效性、透明、问责、知晓滥用、能力性等价值要素。联合国教科文组织与世界科学知识与技术伦理委员会 2016 年 8 月发布了《机器人伦理的报告》，倡导以人为本，努力促使“机器人尊重人类社会的伦理规范，将特定伦理准则编写进机器人中”并且提出机器人的行为及决策过程应全程处于监管之下。国际网络联盟 2017 年 12 月发布《人工智能伦理十大原则》，提出了“系统透明、使用道德黑匣子、服务于人和地球、受人控制、无偏见、福泽全人类、保证公平和自由、建立全球管理机制、遵守法律、禁止军备竞赛”等人工智能伦理准则。

（二）国家层面：各国伦理关注点各有侧重

不同国家的人工智能产业有着不同的发展路径，所呈现的伦理问题关注点亦有差异，因此各国伦理文件各具特色。

1. 美国借助伦理规范保障国家安全

美国基于国家安全的战略高度，强调人工智能伦理对军事、情报和国家竞争力的作用，还发布了全球首份军用人工智能伦理原则，掌握规则解释权。2019 年 2 月，美国总统特朗普发布了《维持美国人工智能领导地位》行政令，重点关注伦理问题，要求美国必须培养公众对人工智能技术的信任和信心，并在应用中保护公民自由、隐私和美国价值观，充分挖掘人工智能技术的潜能。2019 年 6 月，美国国家科学技术理事会发布《国家人工智能研究与发展战略计划》以落实上述行政令，提出人工智能系统必须是值得信赖的，应当通过设计提高公平、透明度和问责制等举措，设计符合伦理道德的人工智能体系。2019 年 10 月，美国国防创新委员会推出《人工智能原则：国防部人工智能应用伦理的若干建议》，对美国国防部在战斗和非战斗场景中设计、开发和应用人工智能技术，提出了“负责、公平、可追踪、可

靠、可控”五大原则。

2. 欧盟推广可信人工智能伦理框架

欧盟认识到，加快发展人工智能技术与其积极推进的数字经济建设密不可分，而要确保数字经济建设长期健康稳定发展，不仅要在技术层面争取领先地位，也需要在规范层面尽早占据领先地位。2019 年 4 月，欧盟高级专家组发布《可信人工智能伦理指南》，提出可信人工智能的概念。专家组从欧洲核心价值“在差异中联合”（united in diversity）出发，指出在快速变化的科技中，信任是社会、社群、经济体以及可持续发展的基石。欧盟认为，只有当一个清晰、全面的，可以用来实现信任的框架被提出时，人类和社群才可能对科技发展及其应用有信心，也只有通过可信人工智能，欧洲公民才能从人工智能中获得符合其基础性价值（如尊重人权、民主和法治）的利益。具体而言，可信人工智能具有三个组成部分：合法性，伦理性和鲁棒性，还包括三层框架：四大基本伦理原则—七项基础要求—可信人工智能评估清单。该框架从抽象的伦理道德和基本权利出发，逐步提出了具体可操作的评估准则和清单，便于企业和监管方进行对照。此外，欧

盟在《人工智能白皮书—通往卓越和信任的欧洲路径》中也提出，赢得人们对数字技术的信任是技术发展的关键。欧盟将创建独特的“信任生态系统”，以欧洲的价值观和人类尊严及隐私保护等基本权利为基础，确保人工智能的发展遵守欧盟规则。

3. 德国率先提出自动驾驶伦理准则

德国依托“工业 4.0”及智能制造领域的优势，在其数字化社会和高科技战略中明确人工智能布局，打造“人工智能德国造”品牌，积极推进自动驾驶领域技术发展。2017 年 6 月，德国联邦交通与数字基础设施部推出全球首套《自动驾驶伦理准则》，提出了自动驾驶汽车的 20 项道德伦理准则，规定当自动驾驶汽车对于事故无可避免时，不得存在任何基于年龄、性别、种族、身体属性或任何其他区别因素的歧视判断，认为两难决策不能被标准化和编程化。

4. 中国倡导发展负责任的人工智能

我国将伦理规范作为促进人工智能发展的重要保证措施，不仅重视人工智能的社会伦理影响，而且通过制定伦理框架和伦理规范，以确保人工智能安全、可靠、可控。为进一步加强人工智能相关法律、

伦理、标准和社会问题研究，新一代人工智能发展规划推进办公室成立新一代人工智能治理专业委员会，2019 年 6 月发布《新一代人工智能治理原则—发展负责任的人工智能》，提出人工智能治理框架和行动指南，强调和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理八项原则。此外，2019 年 8 月，中国人工智能产业发展联盟发布了《人工智能行业自律公约》，旨在树立正确的人工智能发展观，明确人工智能开发利用基本原则和行动指南，从行业组织角度推动人工智能伦理自律。

表 3　各国人工智能伦理文件

国家	文件名	伦理使命	伦理准则	实施细则
美国	《人工智能原则：国防部应用人工智能伦理建议》（2019 年 10 月）	防止战争及确保国家安全，确保安全性和稳健性；符合美国国内法律和美国人民推崇的民主价值	负责、公平、可追踪、可靠、可控	1）在国防部范围内建议工智能指导委员会；2）加强国防部的培训和人才计划；3）投资于研究以增强技术的可重复性的研究；4）加强人工智能测试和评估技术的发展；5）开发风险管理方法；6）确保伦理原则的正确执行等
	《国家人工智能研究与发展战略计划》（2019	人工智能系统必须是值得信赖的	公平、透明、问责	1）通过设计提高公平、透明度和问责制；2）建立道德的人工智能；3）设计符合伦理道德的人工智能体

国家	文件名	伦理使命	伦理准则	实施细则
	年 6 月）			系
	《维持美国人工智能领导地位》（2019 年 2 月）	美国必须培养公众对人工智能技术的信任和信心	保护公民自由、隐私和美国价值观	1）投资人工智能研发；2）制定人工智能技术标准；3）释放人工智能资源
欧盟	《人工智能白皮书—通往卓越和信任的欧洲路径》（2020 年 2 月）	信任的生态系统	以人为本、合乎道德、可持续发展，尊重基本权利和价值观	
	《可信任人工智能伦理指南》（2019 年 4 月）	以人类为中心，旨在服务人类福祉和自由。最大化 AI 系统的利益，最小化其风险：1）合法性；2）伦理性；3）鲁棒性	尊重人类自主性、避免伤害、公平、可解释性原则。具体表现为：1）人的能动性和监督能力；2）安全性；3）隐私数据管理；4）透明度；5）包容性；6）社会福祉；7）问责机制	1. 技术性方法：1）设立可信 AI 的架构；2）将伦理和法律纳入设计；3）解释方法；4）测试和验证；5）服务指标的质量。2. 非技术性方法：1）监管；2）行为守则；3）标准化；4）认证；5）落实责任；6）教育和培养伦理观念；7）利益相关方参与社会对话；8）多元与包容性设计团队
德国	《自动驾驶伦理准则》（2017 年 8 月）	人的生命应该始终优先于财产或动物	保证交通参与者安全、驾驶系统需要官方批准监管、禁止将人群属性作为评价标准、禁止量化生	1）设置监控系统；2）个人数据收集知情同意

国家	文件名	伦理使命	伦理准则	实施细则
			命价值、责任共担	
中国	《新一代人工智能治理原则——发展负责任的人工智能》（2019年6月）	促进新一代人工智能健康发展，人工智能安全可靠可控，更好服务经济发展和社会进步，增进人类共同福祉	和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理	
	《人工智能行业自律公约》（2019年8月）	以人为本、增进福祉、公平公正、避免伤害	安全可控、透明可释、保护隐私、明确责任、多元包容	1）自律自治；2）制定标准；3）促进共享；4）普及教育；5）持续推动

来源：资料整理

（三）企业层面：践行伦理的自我规制之道

面对人工智能引发的负面问题，谷歌、微软、IBM、脸书、旷视、腾讯、百度等科技公司纷纷提出企业层面的人工智能价值观，设立内部的管理机构，践行人工智能伦理原则，以赢得公众的信任。

1. 制定企业伦理原则

科技企业提出了各自的人工智能伦理原则，将其作为规范人工智能的基本框架，为推动人工智能健康发展提供了原则性的指导。IBM公司提出了问责、符合价值、可解释性、公平以及保护用户数据权利等五大企业日常伦理原则，以及打造具备可靠性、公平性、可解释性和可追溯性的四项可信任人工智能原则。谷歌宣布七条伦理原则来指导企业内部的研究和产品开发、商业决策等，同时承诺谷歌将不开发用于武器的人工智能。针对其目的、透明性及技能提出了基本要求微软公司提出人工智能六大原则。我国腾讯、阿里、百度、旷视等行业领军企业也在开展人工智能伦理研究，提出了各自的企业人工智能伦理原则。例如，腾讯倡导面向人工智能的新的技术伦理观，共涵盖三个层面：技术信任，人工智能等新技术需要价值引导，做到可用、可

靠、可知、可控；个体幸福，确保人人都有追求数字福祉、幸福工作的权利；社会可持续，践行“科技向善”。

2. 开展内部机构建设

科技企业在内部机构建设方面，也在扎实开展管理提升工作，加快构建内部控制体系，设立专门机构来切实防范人工智能发展过程中的各项风险。索尼成立名为 Sony AI 的人工智能部门，将通过多个项目及其他探索性研究项目（包括人工智能伦理学）来推动人工智能的基础研究和开发。谷歌宣布成立人工智能道德委员会，即先进技术外部咨询委员会，希望通过哲学家、工程师和政策专家组成的团队帮助解决人工智能带来的伦理风险，但部分委员会成员的身份引发了较大争议，不久之后便予以解散。谷歌旗下专注人工智能的子公司 DeepMind 高度关注人工智能伦理问题，成立了伦理与社会团队（DeepMind Ethics & Society，DMES），DeepMind 还从外部学术机构和慈善机构聘请一些顾问，为该团队提供咨询服务。微软成立了一个人工智能伦理委员会 AETHER，聚集了来自工程、研发、法律、咨询部门的众多专家，旨在为微软的人工智能相关产品和方案制定出

伦理指导原则，用以帮助解决从其人工智能研究、产品和用户互动中产生的伦理和社会、问题。此外，我国旷视科技成立了人工智能创业公司中的第一个人工智能治理委员会—旷视人工智能道德委员会，希望以此推进人工智能应用合理性工作，帮助行业构建一个可持续、负责任、有价值的人工智能生态。百度也分别设立了 AI lab 和百度人工智能体系（AIG），解决人工智能应用场景中的具体问题和人工智能伦理原则的落地方案，打造领先技术平台。

3. 构建企业实践机制

科技企业通过出台人工智能技术实践指南、打造各种使用工具、开展产品评估等多个方面举措，在企业内部践行人工智能伦理的各项规定。IBM 为了落实企业的伦理原则，明确了实现伦理目标的方法论，还推出了各种实用工具。谷歌为明确科学家和工程师们在构建人工智能产品的时候的注意事项，专门出台了一套技术实践指南，还将人工智能治理原则纳入到开发工作，实施审核以及商业协议中，组织多个核心团队来审查隐私保护、歧视等问题，并按产品领域进行评估。为防止算法歧视等问题，微软将人工智能伦理审查纳入即将发布产品

的标准审核清单中。腾讯为了鼓励全社会践行“科技向善”理念，重塑数字社会的信任，呼吁加强科技伦理观的教育宣传和制度化建设，加快研究新兴技术领域的法律规则问题，最终实现技术、人、社会之间的良性互动和发展。百度也提出了要将人工智能隐私保护的价值观融入公司的方方面面，并通过组织、意识、流程制度、技术、产品系统实现隐私保护。

表 4　科技企业人工智能伦理文件

企业	伦理原则	机构建设	企业实践
IBM	1）日常伦理，五大方面 :1. 问责 ;2. 符合价值；3.可解释性；4.公平；5.用户数据权利 2）被信任的人工智能四原则：可靠性，公平性，可解释性，可追溯性	亚马逊，微软，谷歌等企业联合成立了一家非营利性的人工智能合作组织，应对人工智能对伦理和安全带来的挑战	IBM 非常重视可信人工智能，不仅概述实现伦理目标的各种方法论，还推出了各种实用工具
索尼	1）支持创造性生活方式和构建更好的社会；2）利益相关方参与；3）提供可信产品与服务；4）隐私保护；5）尊重公平；6）追求透明；7）持续教育	索尼成立名为 Sony AI 的人工智能部门，以推动人工智能的基础研究和开发	保护客户的个人信息，以及为人工智能的决策机制、基于人工智能的产品和服务的可能影响提供充分的解释

企业	伦理原则	机构建设	企业实践
谷歌	人工智能应：1）有益于社会；2）避免创造或增强偏见;3）为保障安全而建立和测试；4）对人们有说明义务；5）整合隐私设计原则;6）坚持高标准的科学探索；7）根据原则确定合适的应用	谷歌成立了外部顾问性质的人工智能伦理委员会，但由于引发了较大争议，不久之后便予以解散	谷歌出台一套技术实践指南，指导科学家和工程师们在构建产品的时候的注意事项；谷歌将人工智能治理原则纳入到开发工作，实施审核以及商业协议中，组织多个核心团队来审查隐私保护、歧视等问题，并按产品领域进行评估
微软	1）公平： 人工智能应当公平对待每个人；2）包容：人工智能必须赋能每一个人并使人们参与其中；3）可靠性和安全:人工智能必须安全可靠地运行;4）透明；5）隐私与安全; 6）问责：设计、应用人工智能的人员必须对其系统的运行负责	微软成立了内部的人工智能伦理委员会，旨在制定内部政策，确保其人工智能业务和平台符合其核心价值和原则并造福社会	微软将人工智能伦理审查纳入即将发布产品的标准审核清单中，最大努力确保影响人类程序员的隐性偏见不会进入机器学习和人工智能架构
DeepMind	1) 社会福祉；2) 严格与循证；3) 透明与开放；4) 包容与跨学科；5) 合作与包容	DeepMind 宣布成立人工智能伦理与社会部门，目的在于补充、配合人工智能研发和应用活动。但 2019	DeepMind 的三位联合创始人曾签署协议，承诺不会发展致命的人工智能武器系统

企业	伦理原则	机构建设	企业实践
		年 DeepMind 遣散了其人工智能医疗部门的伦理审查委员会	
旷视	确保人工智能正当性、人的监督、技术可靠性和安全性、公平和多样性、问责和及时修正、数据安全与隐私保护	成立了旷视人工智能道德委员会，希望以此推进人工智能应用合理性工作，帮助行业构建一个可持续、负责任、有价值的人工智能生态	将针对采集、传输、存储和使用四个环节的人工智能基础平台形成一套面向数据全生命周期保护，建立一套相关的人工智能数据安全与隐私保护机制
腾讯	1）信任：可用、可靠、可知、可控；2）幸福 :在人机共生的智能社会，确保人人都有追求数字福祉、幸福工作的权利；3）可持续 :践行科技向善，善用技术塑造健康包容可持续的智慧社会	腾讯成立人工智能实验室 AI Lab	1）加强科技伦理的制度化建设；2）加快研究新兴技术领域的法律规则问题；3）加强科技伦理的教育宣传，并鼓励全社会践行“科技向善”理念
百度	1）人工智能的最高原则是安全可控；2）人工智能的创新愿景是促进人类更加平等地获得技术能力；3）人工智能存在的价值是要教人	百度人工智能体系进行组织架构升级，原 AIG（AI 技术平台体系）、TG（基础技术体系）、ACG（百度智能云事业群组）	将隐私保护的价值观融入公司的方方面面，并通过组织、意识、流程制度、技术、产品系统实现隐私保护

企业	伦理原则	机构建设	企业实践
	学习，让人成长，而不是取代人、超越人；4）人工智能的终极理想是为人类带来更多的自由和可能	整体整合为“百度人工智能体系”（AI Group，AIG）	

来源：资料整理

（四）总结：人工智能伦理规制从伦理原则走向伦理体系

人工智能伦理规制之路已经历了两个阶段，第一阶段是人工智能伦理原则的构建，第二阶段是人工智能伦理体系的探索。[2]

1. 第一阶段—已形成共识性伦理原则

迄今已有 40 多个机构或组织提出了各自的人工智能伦理原则，总体上，全球就人工智能伦理原则已基本达成共识，规定呈现趋同化特点，主要包括以下八项原则。

一是人工智能不应取代人—人类福祉原则。人工智能发展应促进社会与人类文明进步，推动自然与社会的可持续发展，也应和平的被利用，避免致命性人工智能武器的军备竞赛。同时，还应创造更加智能的工作方式和生活方式，增进民生福祉。

[2] 陈小平.人工智能伦理体系:基础架构与关键问题[J].智能系统学报,2019,14(04):605-610.

二是人工智能不应伤害人—避免伤害原则。人工智能系统应该保护人类在社会和工作中的尊严、诚信、自由、隐私和安全，不应该增加现有的危害或给个人带来新的危害，同时也应避免对环境和动物造成伤害。

三是人工智能不应分化人—平等均衡原则。人工智能的发展应当尊重各地不同文化习惯，缩小地域差距，鼓励平台、工具、数据、科教等资源的开源开放，实现更好的共享发展，避免将弱势人口置于更为不利的地位。努力破除数据孤岛和平台垄断，不断缩小智能鸿沟，促进人工智能和实体经济深度融合。

四是人工智能不应歧视人—公平正义原则。人工智能系统的开发、使用和管理过程中要保证公平正义，从算法决策、编码设计以及商业应用等不同层面，确保不让特定个人或少数群体遭受偏见和歧视。

五是人工智能不应操纵人—安全可控原则。人工智能系统应不断提升透明性、可解释性、可靠性、可控性，促进公众对人工智能系统的普遍理解，实现算法逻辑、系统决策、行为结果可解释、可预测、

可追溯和可验证。高度关注人工智能系统的安全，提高人工智能鲁棒性及抗干扰性。

六是人工智能不应打扰人—隐私保护原则。人工智能发展应尊重和保护个人隐私，充分保障个人的知情权和选择权，在个人信息的收集、存储、处理、使用等各环节应设置边界，建立规范。同时加强对于个人敏感信息的保护。

七是人工智能不应责难人—责任分担原则。明确人工智能研发、设计、制造、运营和服务等各环节主体的权利义务，在损害发生时，能够及时确定责任主体。倡导相关企业和组织在现有法律框架下创新保险机制，分担人工智能产业发展带来的社会风险。

八是人工智能不应局限人—多元包容原则。促进人工智能系统的包容性、多样性和普惠性。加强跨领域、跨学科、跨国界的合作交流，凝聚人工智能治理共识。力争实现人工智能系统研发人员多元化，训练数据全面化，不因人种、性别、国籍、年龄和宗教信仰等歧视用户。

表 5　人工智能伦理原则、解释说明以及举例

伦理原则		**内涵**	**应用案例（正例或反例）**
人类	**自主**	人工智能应尊重人类意愿	亚马逊智能音箱错误地记

伦理原则		内涵	应用案例（正例或反例）
福祉		与主观能动性，并协助人类进行自主行为	录和理解主人的语音信息，并为主人网购不需要的商品
	自由	人工智能应当促进人类自由地生活、工作、合作等活动	某些人工智能教育类产品，为达到教学目的，限制学生正常的休息与作息
	自律	人工智能应当促进人类活动的自律性，包括遵守法律法规、道德准则等	Google 智能聊天系统 Allo 用戴头巾的人回应持枪表情，存在违法嫌疑
避免伤害	保护个人权益	人工智能应尊重并维护人类个体的各项基本权益	图像处理应用 ZAO 中的“换脸”技术严重侵犯民众的隐私权、肖像权、名誉权等
	确保公共安全	人工智能应当在尊重个人权利的前提下，尊重并维护人类集体的共同的利益	自动驾驶系统既要确保乘客的安全，也要保护行人的安全
	维护国家安全	人工智能应当以促进国家的和谐与安定为根本目的	人工智能不能用于大规模杀伤性武器的开发
平等均衡	均等的教育资源	人工智能应当为公众提供均等的教育资源，保障一致的民众受教育权利	目前人工智能教育及培训主要针对本科及研究生，并未普及大众
	平等的就业机会	人工智能应当保证民众的就业公平，不能剥夺其就业的权利	智能化制造系统导致低技术含量工人下岗；特斯拉无人驾驶出租车导致出租车司机的失业等
	均衡的服务	人工智能应当为公众提供尽量均衡的服务，而非只提供一些特定的服务或应用	目前成熟的人工智能服务主要涉及图像处理、语音识别等，不能提供例如很准确或完善的医疗服务或其他的复杂服务

伦理原则		内涵	应用案例（正例或反例）
	公平的竞争	人工智能应当促进社会公平竞争，维护社会的稳定	人工智能技术有望缩短企业间竞争力差距，促进公平市场竞争
公平正义	**用户非歧视**	人工智能应当对所有用户都非歧视、平等地对待，例如对妇女、儿童、残障人士等特殊人群的非歧视对待	亚马逊的智能招聘系统会给女性较低的评分，有性别歧视嫌疑
	技术非歧视	人工智能中所用到的数据、算法、系统、决策过程等相关技术环节都不应存在歧视用户的行为	电商智能推荐系统会优先推荐其合作伙伴的商品，而非公平推荐
	社会非歧视	人工智能用户不应当歧视非人工智能用户，社会也不应对二者歧视对待	自动驾驶车辆与非自动驾驶车辆应当具有相同的权利和受到一致的法规约束
安全可控	**可审核**	人工智能应当能够在全生命周期内被人类或其他各方审核，包括审核其数据质量、算法效力、设计流程、产品质量等	YouTube 为用户智能推送极端主义/恐怖主义的视频，没有受到审核
	可追溯	应当记录并保留人工智能所用到的数据集、算法、模型，以及开发、部署、维护流程等相关信息，用来追溯并纠正错误	将区块链技术融入人工智能中将确保其可追溯性
	可沟通	人类有权知晓是否正在与人工智能产生互动，并能够通过沟通知晓人工智能的功能及其局限性	自动驾驶系统在一些危险情况时会自主做决策，没有跟驾驶员进行沟通和确认
	可解释	人工智能的技术原理、行	欧盟在其《算法责任与透

伦理原则		内涵	应用案例（正例或反例）
		为目的、行为结果应当能够被清晰、通俗的解释，以增加用户对人工智能的理解程度与信任度	明治理框架》文件中明确定义了算法可解释性，并提出相关技术与非技术实现路径
隐私保护	**必要性原则**	由于模型训练需要采集大量、多样的数据，应当保证数据收集遵循必要性原则，针对使用目的明确数据收集范围	人工智能收集终端可能会收集大量的环境数据或者用户行为日志，例如自动驾驶需要收集大量的环境信息来判断是否是行人
	敏感信息处理	确保计算机视觉、语音识别等应用，过度收集和滥用个人的人脸、声纹的等敏感个人信息	2019 年 8 月，瑞典数据保护机构对当地一所高中开出 20 万元的罚单，理由是学校采用人脸识别系统记录学生的出勤率
	数据安全	确保人工智能系统数据存储安全共享安全和传输安全，防止内部人员盗取数据、数据未经授权访问、数据被泄露等	2019 年 2 月，深网视界被爆泄露超过 250 万人的人脸识别数据
责任分担	**受人监督**	人工智能的所有行为及功能应当可以被人类监控，以确保人类自治	所有涉及到人类生命、财产安全和人类隐私的人工智能应用都应当被监督，如智能医疗、智能交通、自动驾驶、智能金融等人工智能应用
	权限界定	应当明确界定人工智能的哪些行为或功能是能够被允许的，哪些是不允许的，以避免其危害人类权利	在智能医疗中，凡是会对患者产生危害的决定，人工智能不被允许独立执行，需要跟医生进行沟通并最终由医生决定

伦理原则		内涵	应用案例（正例或反例）
	随时接管	人类应有权随时接管人工智能并对其有效控制，以此来避免潜在的风险危害	在自动驾驶中，人类应当做好随时接管的准备
多元包容	**跨学科交流**	鼓励跨学科、跨领域、跨地区、跨国界的交流合作	各国人工智能监管机构人员多元，涵盖计算机、公共政策和哲学等领域专家
	国际合作	开展国际对话与合作，推动形成具有广泛共识的国际人工智能治理原则	OECD 的 36 个成员国联合签署了经合组织人工智能原则

来源：资料整理

2. 第二阶段——逐步探索伦理体系构建

在人工智能伦理原则已经达成共识的基础上，人工智能伦理规制进入了第二阶段，即人工智能伦理体系的构建。因为，任何伦理原则都不可能自我执行，都必须借助于伦理体系中的一系列相互配合的运作机制才能得到落实。



来源：资料整理

图 4　人工智能伦理体系图

人工智能伦理体系包括了三个层次。一是上层的伦理使命（造福人类），人工智能伦理使命虽然也可以作为一条伦理原则，但它本身具有更大的普遍性和稳定性，概括了其他伦理准则的内涵，人工智能的价值在于服务人、造福人，满足人的需要，维持人的生存，促进人的发展，推动人的解放，可以说是其他伦理原则的出发点。二是中

层的伦理原则（共识性的伦理原则），中层的伦理原则即在人工智能伦理规制第一阶段中达成共识的八项伦理原则。三是下层的实施细则（针对具体应用场景，具有可操作的技术性或者非技术性措施），为了将伦理原则落实到一个个具体的应用场景中，解决具体的问题，需要制定一套针对性、可操作的系统化落地方案，这也是伦理体系中非常重要的部分。例如，《欧盟人工智能伦理指南》中为了进一步细化人工智能伦理原则的具体要求，提出了 13 项具体的实施手段（5 项技术手段和 8 项非技术手段）。其中，技术性手段主要侧重人工智能系统的设计与研发阶段，包含建立可信人工智能框架、坚持伦理与法治的设计原则、保证人工智能系统的可解释性、加强对人工智能系统的测试和验证、建立服务质量指标体系五项内容。非技术性手段相对更加关注人工智能系统的使用与监管，主要涉及健全监管法规、出台相应行为准则、设定行业标准、建立资质审核机制、以治理框架实现问责制、通过教育提升伦理意识、确保各方的广泛参与和沟通、维护设计团队的多样性和包容性。

可以看出，人工智能伦理规制不可能由某一领域的专家完成，而是必须涉及一系列相关方，从企业、大学和科研机构、标准化组织、行业组织、政府机构到法律部门，需要所有这些相关方的相互协调和共同努力。在相关方的协调中，不仅需要遵守共同的伦理原则，而且需要遵守伦理体系基础架构对各自角色的定位和相互关系的约定。

四、以法律为保障的风险防控体系

与伦理的性质不同，法律以国家强制力为背书，具有稳定性、强制性、普遍性和滞后性的特征。法律体现统治阶级意志，需要考虑政治、经济、社会等多方面影响，且应当确保技术创新与基本权利保护，以及国家、企业、个人利益之间的平衡。一方面，应考量人工智能所提供的机遇，构建合理的制度环境以激励经济和企业创新发展。另一方面，应防范和应对人工智能技术所带来的各种潜在危害。总体来看，相较于人工智能伦理规范和指引，全球人工智能立法进展较为缓慢，主要体现为数据保护、算法监管等一般性立法对人工智能的源头治理，以及在自动驾驶、金融、医疗等场景化领域中推进立法制定工作。

（一）全球人工智能立法整体趋势

相较以往的科学技术，人工智能所引发的社会风险具有更为鲜明的共生性、时代性和全球性。人工智能技术对当下的法律规则和法律秩序带来一场前所未有的挑战，在著作权法、侵权责任法、道路交通法、劳动法等诸多方面与现有法律制度形成冲突，凸显法律制度供给的缺陷。从整体趋势看，近年全球人工智能的立法监管活动正从慌乱

走向理性，从源头治理走向综合治理，从粗放治理走向精细治理。

1. 数据和算法规制成为人工智能立法的首要命题

从技术维度看，数据、算法和算力是人工智能的三大要素。因此，人工智能的法律问题也主要与其所使用的数据和算法有关，数据的准确性、安全性、隐私性，算法的不透明性、不可控性、问责性均成为人工智能立法首要关注的重要命题。而算力由于其自身纯技术性、客观性的特征，不包含价值问题与伦理问题，因此较少受到立法关注。

数据隐私和数据使用平衡下推进人工智能治理。相关关系而非因果关系、混杂性而非精确性的运行逻辑，意味着海量数据成为人工智能演进优化的根本所在。在此背景下，数据的实时收集、精准分析、规模流动从根本上触动了隐私、公平等人类基本价值。因此，凭借隐私权成熟的理论和实践基础，个人数据保护立法在全球范围内得以迅速推动，并助力人工智能的源头治理。2018 年 5 月，欧盟《通用数据保护条例》（GDPR）正式实施，其中数据的最小化原则、目的限定原则、准确性原则、有限存储原则均对人工智能发展产生直接影响。第 22 条“免受自动化决策权”规定，非经数据主体明确同意，自动

化决策不得使用包括种族、政治观点、宗教、健康数据等在内的敏感数据，借此避免出现根据种族类别分发广告、根据选民政治观点操纵民主进程等问题。以 GDPR 为蓝本，印度、巴西、日本、新加坡、美国加州等国家和地区相继出台或修订个人数据保护立法，通过限制和规范个人数据使用方式以规范人工智能。例如，巴西《通用数据保护法》第六条规定的数据质量原则、透明度原则、不歧视原则对人工智能输入数据的准确性、可控性和非歧视性提出严格要求。

与此同时，OECD、G20、美国、日本等国家和国际组织积极推动跨境数据流动，促进数据向本国回流，为人工智能发展提供原材料。2018 年美国、墨西哥和加拿大签署的《美墨加协定》中加入了高开放程度的跨境数据流动条款。日本在世界经济论坛 2019 年年会上提出了“可信数据自由流动”的概念。2017 年美国《人工智能未来法案》提出促进人工智能领域数据的开源共享和开放。2018 年 12 月，美国国会通过《开放政府数据法案》，要求联邦机构必须以“机器可读”和开放的格式发布任何“非敏感”的政府数据并使用开放许可协议。可见，数据隐私和数据使用利益的平衡已成为人工智能治理的基

本价值取向。

各国加强对生物数据等敏感数据使用的监管指引。相较于互联网对用户上网习惯、消费记录等信息采集，人工智能应用可采集用户人脸、指纹、声纹、虹膜、心跳、基因等具有强个人属性的唯一生物特征信息。因此，除出台一般个人数据保护立法外，各国尤为重视对生物数据等敏感数据的监管和指引。2017 年 5 月，美国华盛顿州通过《关于收集和使用生物识别符的法案》，严格规范企业基于商业用途收集和使用生物识别数据。2019 年 3 月，美国议员参议院提交《商业人脸识别隐私法案》，规定禁止将收集的个人识别数据用于其他目的。2019 年 8 月，瑞典数据保护机构因瑞典一中学违反 GDPR 第 5 条数据最小化原则和第 6 条数据处理的合法性原则，违规利用人工智能技术收集学生生物识别数据，而判处 2 万欧元罚款。2018 年 12 月和 2019 年 8 月，我国分别发布了《信息安全技术 指纹识别系统技术要求》与《信息安全技术 虹膜识别系统技术要求》标准研制，对生物识别系统的数据保护能力提出要求。

透明和问责成为算法专门性立法的核心内容。2016 年 10 月，法

国通过《数字共和法》，尤为强调使用自动化决策的行政机关和数字平台运营者的透明度义务。相较而言，美国强调技术正当程序下的算法问责路径。2017 年 12 月，纽约市通过了《关于政府机构使用自动化决策系统的当地法》，要求成立自动化决策工作组作为实施、监督主体，建立相应的程序机制以确保数据使用的合比例性、解释和获取权、损害救济等问题加以规范。2019 年 2 月，加拿大出台《自动化决策指令》，对算法审查、透明度、质量、用户质疑权加以规定。2019 年 4 月，美参议员提出联邦《算法问责法案》，强调通过诉诸专业性的行政机构或外部监督主体，对算法决策问题进行审查。相较而言，我国算法规制场景过窄，例如《电子商务法》《网络信息内容生态治理规定》《在线旅游经营服务管理暂行规定（征求意见稿）》，均对算法加以规范，但主要集中在新闻推荐、广告推送等定向推送领域，尚未覆盖至医疗、交通、公共安全保障等更为关键的领域。

2. 人工智能立法渐趋理性，更尊重技术规律和法律安定性

有关机器人主体资格的立法讨论和实践热度下降。伴随 2016 年

Alphago 打败人类棋手和 2017 年索菲娅机器人被授予沙特公民身份等焦点事件，有关机器人主体资格的问题引发立法讨论热潮。2017 年 2 月，欧盟议会通过《关于制定机器人民事法律规则的决议》。其中，决议提出应针对更高级的机器人建立登记制度，并引入电子人格以促进自主智能机器人的登记、保险和管理诸多制度。类似地，2017 年，韩国国会和爱沙尼亚政府相继提出人工智能法案，旨在赋予机器人具有相应权利义务的电子人格地位，以厘清潜在的事故责任问题。然而，随着人工智能技术瓶颈和缺陷的显现，科技界的理性发声逐渐被接纳。中国科学院张跋院士指出，“目前全世界的企业界和部分学界对于深度学习技术的判断过于乐观，人工智能迫切需要推动到新的阶段，而这注定将会是一个漫长的过程。”世界各国暂缓立法脚步，截至目前，没有再次出现新的人工智能主体资格立法，相关责任困境转向对算法问责的讨论。同时，欧盟委员会责任与新技术专家组在 2019 年发布《人工智能和其他新兴数字技术责任》报告称，不建议当下赋予人工智能电子人身份，体现出官方认知和态度的转变。

通过解释、修改现有立法规制人工智能，维护法律的稳定性和谦

抑性。2020 年欧盟发布《人工智能白皮书》指出，欧盟有严格的法律框架，如产品责任指令、消费者权益保护法、雇佣和职业平等待遇指令、个人数据保护法，以及健康、运输等特定行业的特别规则等，用以保障人工智能环境下的消费者利益，“这些现行的欧盟法律条款仍将适用于人工智能。”在具体监管中，也体现出对于热点问题或新兴技术风险“回应性治理”的思路，如美国联邦贸易委员会根据《联邦贸易委员会法》第 5 条赋予的广泛执法权，对利用人工智能带来的新型不公平和欺诈性商业行为进行灵活监管。再者，对于人工智能技术难点，通过修改现有立法予以解决。例如，2017 年德国修订《道路交通法》，对自动驾驶的法律含义、责任分配、数据传输等问题作出规定，尝试寻求自动驾驶技术与传统立法的兼容。近期，我国在修改《著作权法》的进程中，也在尝试解决人工智能创作内容的版权保护问题。

对人工智能透明度的要求更加符合产业发展的需要。GDPR 出台时一度引发各国对算法透明度和解释权的追捧，包括美国纽约《关于政府机构使用自动化决策系统的当地法》、加拿大《自动化决策指令》

在内的法案均对算法透明度提出要求。而在实践中，自主学习算法的固有“黑箱”成为法律推动人工智能透明度的现实障碍，同时，人们也更加清晰地认识到透明度对于维护问责、信任价值的局限性，以及对产业发展带来的负面效应。因此，各国立法和监管一方面明晰了对人工智能透明度的实质要求，另一方面注重寻求正当程序、区分规制等其他路径。例如，在 2019 年 4 月欧盟发布的《算法责任与透明治理框架》中指出，算法透明并非对算法的各个步骤、技术原理和实现细节进行解释，算法系统源代码的简单公开也并不能提供有效透明度，反而会威胁数据隐私或影响技术安全应用。美国 2019 年 4 月的《算法问责法案》中也主要强调对企业算法审查的内容，并未提及透明度要求。法国《数字共和法案》也并不要求数字平台运营者履行如行政机关同等的告知义务（包括算法规则、标准以及在最终决策中所占的比重），而仅要求平台为用户提供推荐排名的方式、影响排名的合同关系、第三方报酬等基本信息。

3. 普遍体现以风险为导向的分级分类治理思路

风险成为人工智能立法监管的指向标。风险和技术相伴相生，体

现了技术的不可控性和动态性。自 GDPR 推行“基于风险的路径（risk-based approach）”以来，风险成为各国规制人工智能立法的关键词。为避免过度的法律责任为公私主体，尤其是中小企业带来不合理、不必要的负担，多国立法中强调优先规制“高风险”的人工智能应用。整体来看，高风险的判定一般考虑应用行业、应用场景、覆盖范围等要素。例如，欧盟委员会在《人工智能白皮书》中从三个层面界定了高风险人工智能，首先，结合行业的典型活动特点，人工智能在该行业的应用极有可能引发重大风险，如医疗、运输、能源、移民、司法、救济所等公共行业。其次，行业应用风险的判定应结合具体场景，如在医院预约系统即不属于高风险。第三，基于招聘、监控等特定目的，显著影响个人劳工、隐私等重大利益的，认为属于高风险情形。美国 2019 年 4 月《算法问责法案》同样界定了“高风险自动化决策”的判定标准，同时仅规制年收入超过 5000 万美元或拥有超过 100 万用户的企业。德国则呼吁建立基于风险的五级监管体系，从不受监管的最无害的人工智能系统，到完全禁止最危险的人工智能系统。

风险影响评估逐渐受到各国人工智能立法的认可。风险影响评估

是由技术应用方依据风险大小而作出自适应调节的一种自我规制手段，相比政府监管，具有成本低、灵活性强的特点。GDPR 第 35 条要求数据使用者在应用人工智能自动化决策处理时，应履行数据保护影响评估义务。通过考察人工智能应用中的数据处理活动，评估其使用数据的必要性和合比例性，确定其对自然人权利和自由带来的风险，进而采取相适宜的处理措施。2019 年 2 月，加拿大出台《自动化决策指令》，要求政府部门在生成任何自动化决策系统，或其系统功能、范围发生变化时均需进行影响评估，并通过政府网站发布评估结果。2019 年 4 月，美国国会议员提出的《算法问责法案》，要求对“高风险”的自动决策系统进行影响评估，评估其在准确性、安全、公平、非歧视、隐私等方面的影响。

立法防范公共部门滥用人工智能技术。在美国 Wisconsin v. Loomis 案中，法官使用 COMPAS 软件辅助量刑引发各界对公共部门使用人工智能正当性的关注。相较于商业领域，警务、司法裁判等政府机构或公共领域的智能化应用，通常与人身权利、公正、秩序、非歧视等价值关系重大，体现出更大的应用风险，而既有的公开听证等

权力监督程序如何适用于人工智能技术审查也有待进一步探索。近年，各国率先加紧对公共领域的人工智能应用立法。2017 年 12 月，美国纽约市通过《关于政府机构使用自动化决策系统的当地法》,对法院、警方等公权力机构使用的人工智能自动化决策系统进行安全规制。2019 年以来，美国旧金山、萨默维尔、奥克兰等多个城市禁止政府部门使用人脸识别技术。英国信息专员公署（ICO）发布的《关于执法部门在公共场所使用实时人脸识别技术的建议》则呼吁政府针对实时人脸识别技术的部署和使用，建立强行性的规则和较高的法律门槛，充分平衡各方利益。

（二）人工智能的场景化规制

人工智能的法律规制需要和具体的领域结合起来，算法往往和应用场景、商业模式相结合，在每一个细分领域里，存在着不同的规制方法、进路和手段。各国对于人工智能的法律规制呈现具体化和场景化的特点，在自动驾驶、深度伪造、金融、医疗等具体领域的规制方法、手段、强度和密度存在差异。

1. 助力自动驾驶落地

当前，全球多国已将发展自动驾驶汽车技术上升为国家战略，通过立法加速推进其应用落地。国外的立法较为成熟，许多国家或制定专门的法律，或修订现有的法律，对自动驾驶汽车面临的网络安全、隐私保护、事故责任认定等问题进行了规制，并完善了与自动驾驶汽车配套的保险制度。

美国自动驾驶战略定位明确，政府产业主导与发挥市场机制并行。美国目前已有 30 多个州通过自动驾驶相关立法。自 2016 年起，联邦政府开始出台统一政策，美国交通部相继发布 3 份关于自动驾驶的政策，为自动驾驶的发展提供政策性保障。2017 年 9 月，美国众议院表决通过《自动驾驶法案》，为自动驾驶汽车的监管创建了基本的联邦框架，明确了联邦和州在自动驾驶立法上的职权和分工，避免各州和各政府部门多头管理的局面。2020 年 1 月，美国交通部发布了《确保美国自动驾驶领先地位：自动驾驶汽车 4.0》，确保美国自动驾驶汽车在全球的领先地位。该战略提出涵盖用户、市场以及政府三个方面的十大技术原则，明确了联邦政府在自动驾驶汽车领域的主导地位。

英国着重培育自动驾驶产业环境，促进本国行业竞争力。英国致力于推动自动驾驶技术处于世界领先地位。英国注重个人数据和网络信息安全，在 2017 年 8 月发布的《联网和自动驾驶汽车网络安全关键原则》中提出八大原则，指出评估安全风险，设计和管理安全系统，数据安全存储和传输等的重要性，以保护自动驾驶汽车免于遭到网络攻击的威胁。2018 年 7 月，英国出台《自动化与电动化汽车法》，这是全球首部为自动驾驶设计保险制度的法律，明确自动驾驶汽车发生事故后，可根据车辆的投保情况由保险公司以及车主来承担事故损失带来的赔偿责任。此外，英国还启动了自动驾驶汽车法律审查机制，针对自动驾驶事故发生责任确定、自动驾驶汽车刑事犯罪等新议题革新现有法律，推动自动驾驶法律跟上技术发展的步伐，为英国率先在高速公路上使用自动驾驶汽车铺路。

日本立法环境较为开放，明确自动驾驶技术推进时间表。日本是目前为止自动驾驶相关法律方面走在世界最前端的国家之一，鼓励高级别的远程无人自动驾驶的测试，并对路测主体设置了更为全面的安全义务性规定，提出到 2020 年在一定条件下实现在高速公路和人口

稀少地区自动驾驶的目标。2018 年 3 月，日本政府提出《自动驾驶相关制度整备大纲》，对自动驾驶汽车发生事故时的责任问题加以明确，规定自动驾驶事故损失继续使用《机动车损害赔偿保障法》中对民事责任的要求。2019 年 5 月，日本通过了《道路运输车辆法》的修正案，为实现自动驾驶产业化规定了安全标准。

韩国立法推动自动驾驶商业化，率先提出自动驾驶安全标准。近年来，韩国政府大力发展智能汽车，将自动驾驶纳入国家战略，提出 2030 年成为未来型汽车世界强国的目标。2019 年 4 月，韩国修订《汽车事故赔偿法》，规定 L3 级要求驾驶员应当随时准备接管自动驾驶汽车，出现事故的，主要责任人仍是驾驶者本人。此前，韩国政府出台《自动驾驶汽车商用化促进法》，规范自动驾驶汽车提供商业服务的行为。可见，韩国政府正在为“成为全球第一个将自动驾驶商业化的国家”加紧制定监管和法律框架。2020 年 1 月，韩国国土交通部发布《自动驾驶汽车安全标准》，针对自动驾驶汽车的部分功能提出有条件自动驾驶车（L3 级）安全标准，为驾驶员提供安全保障。

我国在立法推进上更为谨慎，推动形成智能网联汽车发展路线。我国在自动驾驶的正式立法尚在研究当中，目前仍然以有关部门的政策指导的方式来实行自动驾驶汽车的监管和规范。2018 年 4 月，工

信部、公安部、交通运输部联合发布了《智能网联汽车道路测试管理规范（试行）》，对测试主体、测试驾驶人及测试车辆、测试申请及审核、测试管理、交通违法和事故处理等方面作出规定。2020 年 2 月，发改委、网信办、工信部等 11 个部委联合印发了《智能汽车创新发展战略》，提出到 2025 年，中国标准智能汽车的技术创新、产业生态、基础设施、法规标准、产品监管和网络安全体系应基本形成。

表 6　各国对于自动驾驶的立法规制

国家	文件名	发布时间	主要内容
美国	《准备迎接未来的交通：自动驾驶汽车 3.0》	2018 年 10 月	该文件致力于推动自动驾驶技术与地面交通系统多种运输模式的安全融合
	《自动驾驶法案》	2017 年 9 月	该法案旨在促进自动驾驶技术和汽车发展，规定美国联邦对自动驾驶汽车设计、制造和性能的立法优先权，允许自动驾驶汽车在公共道路上测试，显著增加自动驾驶汽车豁免的数量并逐年提高，成立自动驾驶汽车委员会探索自动驾驶汽车安全标准
	《联邦自动驾驶汽车政策》	2016 年 9 月	该政策围绕自动驾驶汽车性能指南、统一的州政策、现行国家公路交通安全管理局(NHTSA)监管工具、新的监管工具展开，为自动驾驶安全部署提供政策监管框架，从而为有效利用技术变革提供指导意见
英国	《自动与电动	2018 年 7 月	该法对于自动驾驶汽车的保险制度问题

国家	文件名	发布时间	主要内容
	汽车法》		做了专门规定
	《联网和自动驾驶汽车网络安全关键原则》	2017 年 8 月	八大原则分别为：董事会部署、管理并改进组织机构安全；适当、按比例评估、管理安全风险，包括供应链特有的安全风险；组织机构需部署产品后期维护和事件响应，确保系统在整个生命周期的安全；所有组织机构，包括子承包商、供应商和第三方应合作改进系统安全；采用深度防御方式设计系统；在软件整个生命周期内，对所有软件进行管理；确保数据存储与传输安全、可控；确保系统对攻击具有弹性，当防御机制或传感器失效时，确保能适当予以响应
日本	《道路运输车辆法》	2019 年 5 月	该法为实现自动驾驶实用化规定了安全标准，允许 L3 级自动驾驶汽车在公路上行驶
	《自动驾驶相关制度整备大纲》	2018 年 3 月	该大纲规定自主行驶时的事故赔偿责任原则上由车辆所有者承担，可以利用交强险进行赔付
韩国	《自动驾驶汽车安全标准》	2020 年 1 月	该标准对自动驾驶汽车的部分功能提出有条件自动驾驶车（L3 级）安全标准
	《汽车事故赔偿法（修订）》	2019 年 4 月	该法明确涉及 L3 级自动驾驶汽车的事故责任。由于 L3 级要求驾驶员应当随时准备接管自动驾驶汽车，因此若涉及 L3 级自动驾驶汽车的事故，主要责任人仍是驾驶者本人
	《自动驾驶汽车商用化促进法》	2019 年 4 月	该法规定了自动驾驶汽车商用化服务规范措施，自 2020 年 5 月 1 日起正式实施
	《韩国汽车管	2017 年 2 月	该法允许在城市道路上测试自动驾驶汽

国家	文件名	发布时间	主要内容
	理法》		车，这意味着韩国道路将成为自动驾驶汽车测试场
中国	《智能汽车创新创新发展战略》	2020年2月	战略提出到2025年，中国标准智能汽车的技术创新、产业动态、基础设施、法规标准、产品监管和网络安全体系基本形成
	《智能网联汽车道路测试管理规范（试行）》	2018年4月	该规范主要明确了规范中明确了测试主体、测试驾驶人及测试车辆应具备的条件，以及测试申请及审核，测试管理，交通违法和事故处理等内容
	《国家车联网产业标准体系建设指南（智能网联汽车）》	2017年12月	该指南主要针对智能网联汽车通用规范、核心技术与关键产品应用，指导车联网产业智能网联汽车标准化工作，加快构建包括整车及关键系统部件功能安全和信息安全在内的智能网联汽车标准体系

来源：资料整理

2. 防止深度伪造滥用

美国专门针对深度伪造予以立法规制，欧盟、德国、新加坡则将深度伪造涵盖在不实信息或虚假新闻的规制空间内。但不管是采取哪种模式，其目的都是为了最大限度地降低深度伪造技术对个人、社会及国家带来的危害。

美国直接对深度伪造进行立法，联邦和各州政府层面均有部署行动。美国担心深度伪造对2020年大选和国家安全的影响，2019年美

国国会先后提出了《2019 年深度伪造报告法案》和《深度伪造责任法案》。此外，美国的加州、德州、马萨诸塞州、弗吉尼亚州等也陆续推出了相关立法。这些立法提出的主要措施包括：一是划定应用边界，禁止政治干扰、色情报复、假冒身份等非法目的的深度伪造，否则可能构成刑事犯罪；二是设置披露义务，要求制作者、上传者以适当方式披露、标记合成内容，例如采取嵌入数字水印、文字、语音标识等方式；三是加强技术攻防，呼吁开发检测识别技术和反制技术。

欧盟主要通过虚假信息治理等限制深度伪造技术的应用。欧盟委员会于 2018 年 4 月发布的《应对线上虚假信息：欧洲方案》集中阐释了欧盟委员会面对线上虚假信息挑战的基本观点，提出改进信息来源及其生产、传播、定向投放和获得赞助方式的透明度要求，还规定了改善信息的多样性，提高信息的可信度，制定包容性的解决方案等原则，以实现全面防范视频、图像和文字等虚假信息，避免信息发布者违法操纵舆论等状况。2018 年 9 月，欧盟发布其历史上首份《反虚假信息行为准则》，旨在加强互联网企业对平台内容的自我审查，从源头打击网络虚假内容。

德国利用现有规制网络内容立法，防范深度伪造技术的危害。以深度伪造为代表的人工智能造假技术，可以被德国已有互联网内容治理法律所调整。例如德国 1997 年出台的《信息和传播服务法》和 2018 年通过的《社交媒体管理法》，要求社交媒体公司检查被投诉内容，限制互联网提供商传播非法内容，设置“网络警察”监控危害性内容传播，并对制作或传播对儿童有害内容的言论视为犯罪行为。除此以外，德国相继出台的与网络舆情、虚假信息治理相关的法律法规，都可以用来有效打击人工智能深度伪造行为。

新加坡针对网络虚假内容立法，兼顾调整深度伪造行为。2019 年 5 月，新加坡通过《防止网络虚假信息和网络操纵法》，该法适用于利用深度伪造技术制作的虚假音视频，规定政府有权要求网络平台或个人删除损害公共利益的虚假信息，并对不遵守指示的网络平台或个人判处监禁或并处罚金。

我国相关立法也开始关注深度伪造问题，侧重人格权保护和内容管理。我国《民法典》和《网络信息内容生态治理规定》中对深度伪造问题进行了回应，规定禁止利用信息技术手段伪造的方式侵犯他人

的肖像权和声音；不得利用深度学习、虚拟现实等新技术新应用从事法律、行政法规禁止的活动。此外，《网络音视频信息服务管理规定》还专门对深度伪造问题进行了一系列的制度设计，包括规定安全评估要求、显著方式予标识义务、不得制作、发布、传播虚假新闻信息、建立健全辟谣机制等。

表 7 各国对于深度伪造的立法规制

国家	文件名	发布时间	主要内容
美国	《2019 年深度伪造报告法案》	2019 年 6 月	该法案明确了“数字内容伪造”的定义，规定国土安全部定期发布深度伪造技术相关报告
	《深度伪造责任法案》	2019 年 6 月	该法案要求任何创建深度伪造视频媒体文件的人，必须用“不可删除的数字水印以及文本描述”来说明该媒体文件是篡改或生成的，否则将属于犯罪行为
	《关于制作欺骗性视频意图影响选举结果的刑事犯罪法》	2019 年 6 月	该法将利用 Deepfake 等技术制作深度伪造视频企图干扰选举的行为定义为刑事犯罪
	《2018 年恶意伪造禁令法案》	2018 年 12 月	该法对制作深度伪造内容引发犯罪和侵权行为的个人，以及明知内容为深度伪造还继续分发的社交媒体平台，进行罚款和长达两年的监禁。如果伪造内容煽动暴力、扰乱政府或选举，并造成严重后果的，监禁将长达 10 年

国家	文件名	发布时间	主要内容
欧盟	《反虚假信息行为准则》	2018 年 9 月	该准则旨在加强互联网企业对平台内容的自我审查，从源头打击网络虚假内容
	《应对线上虚假信息：欧洲方案》	2018 年 4 月	该方案集中阐释了欧盟委员会面对线上虚假信息挑战的基本观点，提出改进信息来源及其生产、传播、定向投放和获得赞助方式的透明度，改善信息的多样性，提高信息的可信度，制定包容性的解决方案等原则，以实现全面防范视频、图像和文字等虚假信息，避免信息发布者违法操纵舆论等状况
德国	《社交媒体管理法》	2018 年 1 月	该法要求社交媒体公司必须设立有关程序，检查自己网站上被提出投诉的内容，并在 24 小时之内删除明显违法的信息
	《信息和传播服务法》	1997 年 6 月	该法涉及互联网服务商的责任、保护个人隐私、数字签名、网络犯罪和保护未成年人等方面，是一部全面的综合性的法律
新加坡	《防止网络虚假信息和网络操纵法》	2019 年 5 月	该法使政府有权要求个人或网络平台更正或撤下对公共利益造成负面影响的虚假内容，该法适用于利用深度伪造技术制作的虚假音视频
中国	《民法典》	2020 年 5 月	该法规定不得利用信息技术手段伪造等方式侵害他人的肖像权
	《网络信息内容生态治理规定》	2020 年 3 月	该法规定不得利用深度学习、虚拟现实等新技术新应用从事法

国家	文件名	发布时间	主要内容
			律、行政法规禁止的活动
	《网络音视频信息服务管理规定》	2019 年 11 月	该规定对网络音视频服务使用者和提供者均提出要求，即利用基于深度学习、虚拟现实等的新技术应用制作、发布、传播非真实音视频信息的，应当以显著方式予以标识，不得利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播虚假新闻息，网络音视频信息服务提供者应当建立健全辟谣机制

来源：资料整理

3. 规范智能金融产品

金融是现代经济的核心，金融服务行业也一直是技术创新的积极实践者和受益者。人工智能目前正广泛应用到金融业中，形成了智能风控、智能投资、智能交易、智能投顾等应用场景。具体而言，智能金融是指，以人工智能为代表的新技术与金融服务、产品的深度融合。智能金融需要有新的监管技术，各国对智能金融产品设置了灵活性的监管规定。

美国直接在立法文件上明确人工智能投资产品的法律义务。2017 年 2 月，美国证券交易委员会（SEC）的投资管理部发布了《智能投

顾指南》，强调了机器人顾问在履行《顾问法》所规定的法律义务。同一时期，SEC 的投资者教育与倡导办公室还发布了《投资者公告》，该公告旨在教育个人投资者有关机器人顾问的知识，并帮助他们决定是否使用机器人顾问来实现其投资目标。

欧盟通过一般法令强化对于人工智能金融产品对监管。欧盟发布了《网络和信息系统安全指令》指出，在银行和金融市场基础设施领域，操作风险是审慎监管的重要组成部分。《不公平商业指令行为》中规定，禁止不公平的商业行为，列明违背专业勤勉要求，并有可能曲解消费者的经济行为。其中，具有误导性的行为、遗漏行为以及侵略性行为（包括通过电话、电子邮件或其他媒体进行持续不断的不必要的推广）都属于不公平的商业行为。

我国开始对人工智能金融应用进行立法，侧重于鼓励行业发展和防范风险。2017 年 7 月，国务院发布的《新一代人工智能发展规划》提出了“智能金融”的概念，明确指出要建立金融多媒体数据处理与理解能力，创新智能金融产品和服务，发展金融新业态。2018 年 4 月，中国人民银行、中国银行保险监督管理委员会、中国证券监督管

理委员会、国家外汇管理局印发了《关于规范金融机构资产管理业务的指导意见》，对人工智能在金融领域的应用进行了规制，从胜任性要求、投资者适当性以及透明披露方面对智能投顾中的算法进行穿透式监管。《证券法》规定，通过计算机程序自动生成或下达交易指令进行程序化交易的，应当符合国务院证券监督管理机构的规定，并向证券交易所报告，不得影响证券交易所系统安全或者正常交易秩序。

表 8　各国对于智能金融产品的规制

国家	文件名	发布时间	主要内容
美国	《智能投顾南》	2017 年 2 月	指南重点关注以下三个不同的领域，并就机器人顾问如何应对这些问题提出建议：1)在向客户披露机器人顾问及其投资咨询服务的相关情况时，披露的内容及采取的陈述方式；2)为了给客户提供适当的建议，从客户处收集相应信息的责任；3)采取并实施有效的合规管理制度，合理设计制度内容，以解决与自动化提供建议相关的特定问题
	《投资者公告》	2017 年 2 月	该公告旨在教育个人投资者有关机器人顾问的知识，并帮助他们决定是否使用机器人顾问来实现其投资目标
欧盟	《网络和信息系统安全指令》	2016 年 7 月	该立法指出在银行和金融市场基础设施领域，操作风险是审慎监管

国家	文件名	发布时间	主要内容
			的重要组成部分
	《不公平商业指令行为》	2005 年 5 月	该立法禁止不公平的商业行为，列明违背专业勤勉要求，并有可能曲解消费者的经济行为
中国	《证券法》	2019 年 12 月	该立法涉及到智能金融监管的条款主要有两条：关于程序化交易的规制和对于不以成交为目的的恶意申报与撤单的规制
	《新一代人工智能发展规划》	2017 年 7 月	该规划强调要深化金融体制改革，健全金融监管体系
	《关于规范金融机构资产管理业务的指导意见》	2018 年 4 月	该意见对智能投顾资质、算法备案管理、程序化交易算法失效的监管做出规定

来源：资料整理

4. 促进智能医疗发展

近年来，人工智能技术与医疗健康领域的融合不断加深。人工智能技术也逐渐成为影响医疗行业发展，提升医疗服务水平的重要因素。智能医疗的应用场景主要包括：语音录入病历、医疗影像辅助诊断、药物研发、医疗机器人等方面。目前美国、欧盟和我国纷纷出台文件促进人工智能医疗的发展。

美国发布相关指南，针对智能医疗进行广泛指导。2019 年 9 月，美国食品药品管理局（FDA）发布《器械软件功能和移动医疗应用政

策指南》，以告知制造商、分销商和其他组织 FDA 如何监管移动平台或通用计算平台上使用的软件应用程序。2019 年 9 月，FDA 还颁布了《临床决策支持指南草案》，阐释临床决策支持软件的定义以及 FDA 对其监管的范围。

欧盟通过立法为智能医疗建立起体系化的监管方案。欧盟于 2017 年 4 月发布了《欧盟医疗器械法》，该立法在器械的分类、器械的通用安全和性能要求、技术文件的要求及上市后监管体系方面进行了相应改变，细化了多条性能要求，强调将风险分析和管理贯穿于设计和生产、销售、上市后监管整个产品周期中，并要求建立警戒和上市后监管电子系统。

我国基于鼓励发展的态度，开展对于智能医疗的精细化监管。2017 年 2 月，国家卫计委发布《人工智能辅助诊断技术管理规范(试行)》以及《人工智能辅助治疗技术管理规范(试行)》，对使用计算机辅助诊断软件及临床决策支持系统提出要求，明确以诊断准确率、信息采集准确率、人工智能辅助诊断平均时间以及人工智能辅助诊断增益率作为人工智能辅助诊断技术临床应用的主要考核指标，为人工智

能应用于临床诊断和治疗提供了规范。2018 年 5 月，国务院发布《关于促进“互联网+医疗健康”发展的意见》，提出要推进“互联网+”人工智能应用服务，研发基于人工智能的临床诊疗决策支持系统，开展基于人工智能技术、医疗健康智能设备的移动医疗示范。

表 9　各国对于智能医疗的规制

国家	文件名	发布时间	主要内容
美国	《器械软件功能和移动医疗应用政策指南》	2019 年 9 月	指南告知制造商、分销商和其他组织如何监管移动平台或通用计算平台上使用的软件应用程序
	《临床决策支持指南》	2019 年 9 月	指南进一步阐释临床决策支持软件的定义以及 FDA 对其监管的范围。还新增加了对患者决策支持软件的讨论，将其与临床决策支持软件并列分析
欧盟	《欧盟医疗器械法》	2017 年 4 月	该立法在器械的分类、器械的通用安全和性能要求、技术文件的要求及上市后监管体系方面进行了相应改变
中国	《关于促进“互联网+医疗健康”发展的意见》	2018 年 5 月	该意见提出要推进“互联网+”人工智能应用服务，研发基于人工智能的临床诊疗决策支持系统，开展基于人工智能技术、医疗健康智能设备的移动医疗示范
	《人工智能辅助诊断技术管理规范(试行)》	2017 年 2 月	该规范主要对使用计算机辅助诊断软件及临床决策支持系统提出要求
	《人工智能辅助治疗技术管理规范(试行)》	2017 年 2 月	该规范专门对使用机器人手术系统辅助实施手术的技术提出要求

来源：资料整理

五、人工智能治理展望

人工智能的技术发展与产业应用是全球高度协作的成果，其未来的发展依赖世界各国优势互补、合作共享，而人工智能带来的社会风险也具有共生性、时代性、全球性的特点，没有一个国家能够独善其身，需要在全球范围内达成治理共识。因此，围绕全球人工智能治理的共同问题，需要立足未来，加强国际合作，构建全球人工智能治理的命运共同体。2019 年 11 月，在联合国教科文组织（UNESCO）召开的第 40 届大会期间，193 个会员国决定委托该组织就人工智能伦理问题制定第一份全球规范性文件，重点关注人工智能对公平正义和人类权利带来的挑战，强调以符合伦理的方式开发和使用人工智能，并支持推动可持续发展目标方面的国际合作。[3]

（一）坚持包容、弹性的治理理念

对于人工智能治理问题的纾解方式，应放在不同国家的人工智能发展路径下，以及不同的场景、业务下进行讨论，采取因事制宜，包容、弹性的治理理念。一方面，不同国家在解决人工智能治理相关问

[3] 参见 UNESCO C40/37 号决议，2019 年 11 月 26 日根据社会科学及人文科学委员会的报告通过。

题（如隐私保护、算法透明、数据跨境流动等）时，由于具有不同的文化历史背景和产业发展基础，拥有不同的治理目的和价值，实行不同的法律和监管制度，因此所采取的治理方式可能存在一定的差异。以隐私保护为例，虽然各国都认同应该保护隐私，但不同国家对于隐私的理解，以及隐私的重要性排序有着明显差异，存在“隐私优先”（High-Privacy position）、“隐私平衡”（Balance-Privacy position）和“隐私限制”（Limited-Privacy position）不同的价值判断。因此，在现阶段开展的国际人工智能治理政策讨论和制定过程中，既需要突出全球各国的共同理想和追求，例如公平非歧视、最大化福祉等方面，也需要尊重和理解各国制度设计中存在的不同，才能在合作中寻求到良好的治理效果。另一方面，人工智能的发展具有不确定性，既有新的治理问题，也有老的监管问题，有些治理问题是传统法律调整后可以解决的，但一些新的监管问题，难以适用统一的规则或标准予以约束，应当从场景和领域出发进行具体的分析和讨论。根据人工智能使用场景、影响范围、可能的危害程度的不同，应当采用分类治理的思路。对涉及国家安全、人民群众生命财产安全、社会稳定重大敏感利

益等“高风险”的人工智能应用场景，例如自动驾驶、智能医疗等领域，应当加强事前监管，按合理审慎的原则设置必要的准入限制和约束条件；对于与个人消费及服务相关的电子商务、物流、智能家居等“低风险”人工智能应用领域，采取基于结果的规制思路，侧重于事中事后的监管。

（二）构建不同阶段的治理路径

人工智能是人类社会的伟大发明，同时也存在巨大的社会风险，包括“技术—经济”决策导致的风险，或者是法律保护的科技文明本身带来的风险。法律与伦理作为两种重要的调整手段，可以通过不同方面、以不同方式、采用不同机制对社会生活的不同领域发挥各自的影响和作用。由于人工智能技术的特殊性，既需要以法律规则为代表的强制性要求，规制与国家安全、社会公共利益和个人权益密切相关的治理问题，也需要以伦理指南为代表的规范性要求，引导产业、企业进一步履行社会责任，增强产品和服务的安全性、可靠性和鲁棒性。在法律和伦理相互配合的基础上，可以通过区分人工智能各发展阶段，来明确不同的治理重点和目标。

1. 近期应加快制定产品和服务标准，利用“数据治理”推动人工智能治理问题的解决

这一阶段仍为“弱人工智能”，即人类主体能够以某种方式对人工智能产品、服务进行干预，避免其因为设计缺陷、使用不当，对社会和个人带来负面影响。也就是说，即使人工智能治理问题出现，某一方主体依然可以通过某种方式解释相关决策的不当，或通过经济手段弥补带来的过失。因此，现阶段的法律问题更多体现在传统法律适用，而不是创设新的人工智能法律制度，如《自动化决策法》《人工智能法》等。反而应突出技术治理发挥的作用，聚焦于影响人工智能决策结果的要素规制，最大程度地控制和减少人工智能带来的公共危险和对私人权利的侵犯。

在治理重心方面，本阶段应重点关注人工智能应用数据对个人权益、国家安全和企业利益带来的影响，以及人工智能在重点行业领域的应用。一方面，应借助“数据治理”实现对于算法和应用的治理，加快个人信息保护、数据安全、数据跨境流动、数据共享交换等制度建设，平衡数据安全与数据价值之间的关系，充分挖掘数据潜在的价值，同时尽可能降低数据利用的成本和控制可能产生的风险。鉴于数

据多元主体的现状，需要构建“多方参与、分层监管、合理担责”的治理体系，协调政府、企业和用户在数据使用中的关系。同时还应将宏观数据治理规则精确提现到具体的应用场景，充分发挥数据的作用，挖掘数据价值。另一方面，明确的行业规范是人工智能得以创新发展的前提。交通、医疗、金融、新闻等领域与个人生命健康、财产安全、社会稳定密切相关，可以借由行业协会或产业联盟对市场进行规范，以出台行业标准规范的方式提升产品和服务的稳定度与质量，同时辅以保险、承诺等手段弥补可能对个人或社会带来的损失。

在治理方式上，企业负责产品的研发与运营，将是本阶段主要的治理主体，可以依托平台、技术等手段实现自我监督，政府监管部门并不需要过多干预产品和服务上线前的流程。针对人工智能的风险，企业需要有针对性的准备措施及预案，确保人工智能系统在其整个生命周期内安全、可靠、可控地运行。企业应当加强评估人工智能系统的安全性和潜在风险，不断提高系统的成熟度、稳健性和抗干扰能力，确保系统可被人类监督和及时接管，避免系统失控的负面影响。当产品和服务进入市场后，政府部门应结合监管需求，对数据市场竞争、

个人信息保护、数据安全制度执行情况进行监管。



2. 中远期应调整责任法律制度，实现法律与伦理相衔接

这一阶段的特点是人工智能的自主性将进一步增强，可能脱离了设计者的目标和初衷，但人类仍可以通过某种方式对人工智能系统输出结果进行干预，但应遵循“非必要不干预”原则。在这一阶段下，人工干预与机器自主决策同时发挥作用，对于法律制度的影响更多将体现在责任机制的调整上，尤其是传统的产品责任能否适用于人工智能为代表的软件及系统服务。2020 年 2 月 19 日，欧盟委员会发布的《关于人工智能、物联网和机器人对安全和责任的影响的报告》提及了这一重要问题，进一步分析了人工智能、物联网和其他数字技术对欧盟相关安全和责任立法框架的影响，提出了包括人工干预辅助、强制算法透明度要求、供应链法律责任分配、联合产品安全立法等建议，为讨论下一阶段人工智能治理重点和方式提供了有益参考。

在治理重心上，由于对于产业链上的不同主体实现公平有效的责任分配可能是很困难的，以伦理、治理指南等为代表的规制方式将是此阶段治理的中心。可以通过出台涉及国家、行业、企业的治理指南，出台标准指引等方式加强相关指南原则的落地实施，避免人工智能系

统用于非法或违反伦理的目的。与此同时，各国应重新评估自发性的人工智能技术及应用对现行侵权责任、刑事责任体系带来的影响，明确人工智能研发、设计、制造、运营和服务等各环节主体的权利义务，研究在产品责任法律体系中纳入软件等新要素的可行性，同时在部分领域探索算法透明度的要求和程序。

在治理方式上，政府应倡导相关企业和组织在法律框架下创建责任分担机制，合理分担人工智能可能带来的风险。企业应加强社会宣传教育，提升人工智能用户人群的自我权益保护意识，降低人工智能发展过程中可能存在的伦理风险。第三方行业协会及产业联盟将是政府、企业之外更加重要的治理主体。一方面，第三方机构可以更好体现产业发展的诉求，牵头撰写相关的治理原则及落地措施，避免强制性规则对产业发展带来不必要的阻碍。另一方面，在涉及政府部门对于算法等新技术要素监管方面，第三方机构能够发挥专业作用，组织开展相关的评估和测评工作，为监管提供有效的支持和帮助。此外，国际组织也需要提前考虑超人工智能时代对于人类权利和社会体系的影响，明确未来社会中的人类地位与基本权利，实现可持续发展的

目标。

（三）打造多元共治的治理机制

人工智能治理的重要特征之一是治理主体的多元化、治理手段的多样化，需要打造多元主体参与、多措并举、协同共治的治理机制。这依赖于包括国际组织、政府、行业、企业、公众等多利益攸关方的参与合作，需要各方各司其职、各尽其能，以适当的角色、最佳的手段协同共治。可以说，人工智能这样的新技术正在催生社会治理体系和机制的变革，从传统的政府主导的权威式自上而下的单向管理走向多元主体协同共治的治理新范式，综合多个主体、多种手段的优势，保持开放灵活的状态，及时根据技术发展需要进行动态调整。

从国际层面来看，政府间国际组织可以通过开展国际对话、协调与合作的方式，确保人工智能技术的发展和应用能够造福人类。国际合作并不意味着各国一定要遵循同一套人工智能规范、标准或者法规，也不意味着方方面面至始至终需要订立国际协定。[4]而是在充分尊重各国人工智能治理原则和实践的前提下，探索各国在人工智能规范及

[4] ÓhÉigeartaigh, S.S., Whittlestone, J., Liu, Y. et al. Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance. Philos. Technol. (2020).

指导方针方面重叠共识的领域，从而确保人工智能对人类有益。可采取的合作方式包括：各国人工智能研究人员合作完成项目，确保人工智能安全可控；建立各国平等沟通的渠道，确保国际讨论能够汲取多样的国际视角；邀请各国利益相关者寻求共识、弥合分歧，参与制定治理规则等。

从国家层面来看，政府应当肩负起对人工智能技术及服务监督管理的职责，为人工智能营造良好的发展环境。首先，人工智能作为一项集合多学科知识、高度复杂的技术，政府应当优化相关职能机构，选拔具有人工智能背景的专业人才参与到政府治理过程中，为人工智能治理体系的构建提供理论支撑，从而以务实审慎的态度推动人工智能的发展。其次，人工智能所带来的创新多样性和不可预测性，需要更具弹性、适应性和可操作性的监管方式。政府应当从基于具体规则的监管方式转变向基于原则的监管方式，避免过于详细、严格的事前监管，可以将监管细节留给人工智能指南、行为规范、认证规则等进行更加灵活的规定。同时，政府应加强与行业、企业的互动和交流，增强监管决策的针对性和有效性，避免在对人工智能主观预测的基础

上做出决策。最后，政府应当建立健全配套体系，可以在先行试验基础上评估人工智能的社会影响，适时修订立法中与人工智能发展不相适应的部分，探索可能的算法监管路径，防范数据安全风险，构建结构合理且责任明晰的人工智能产品责任体系。

从行业组织层面来看，行业组织可以通过制定技术标准、倡导行业自律等方式推动人工智能规范发展。人工智能治理需深入行业中去，从每个行业的实践中去探索和实践。一方面，行业组织可以推动构建动态的人工智能研发应用评估评价机制，围绕人工智能设计、产品和系统的复杂性、风险性、不确定性和可解释性等问题，制定测试方法和标准体系，推动人工智能安全认证。另一方面，行业组织可以制定行业自律公约，明确人工智能开发应用的基本原则和行动指南，为相关企业提供行业自律指导和建议，强化企业的社会责任意识，推动行业自律有效开展。

从企业层面来看，企业可以通过建立内部自律机制、研发技术工具等方式，推动治理举措的落地实践。一方面，交通、医疗、金融等行业的相关领军企业在人工智能研发和应用过程中应当强化社会责

任意识，推动行业自治，制定人工智能行业从业人员行为规范，加强从业人员自我约束，开展伦理审查，确保企业各环节严格遵守伦理规则。另一方面，企业加快推动人工智能安全可信技术研发，强化人工智能产品和系统的网络安全防护能力，全球企业的交流协作也至关重要，广泛的交流有助于解决人类共同面临的人工智能技术创新和探索。

从公众层面来看，公众是治理过程中的重要参与者，可以适当介入到治理监督过程中，维护自身合法权益。一方面，公众可以通过人工智能产品的投诉举报等监督渠道，实现对人工智能技术的监督和意见反馈，为人工智能治理献计献策，使人工智能发展真正地“以人为本”。另一方面，公众可以参加人工智能公共伦理教育，减轻对于人工智能技术的恐慌，提高自身的伦理安全防范意识。公众还可以积极参与岗位被替代人员的数字劳动技能再培训活动，主动应对现有和未来的劳动力挑战问题。

人工智能时代已经到来，各方主体要以伦理的力量、法律的理性引领和规制人工智能技术的发展，确保人工智能技术更加安全可控，

更合乎伦理和法理，使人工智能成为促进社会有序发展、共享发展、公平发展、开放发展、和谐发展的生产力基础。[5]



[5] 张文显.构建智能社会的法律秩序[J/OL].东方法学:1-14[2020-09-09].

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010-62304839

传真：010-62304980

网址：www.caict.ac.cn

中国人工智能产业发展联盟

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010-62302973

传真：010-62304980

网址：www.aiiaorg.cn

